United States
Environmental Protection
Agency
Air

Office of Air Quality
Planning and Standards
Research Triangle Park, NC 27711

EPA-454/R-01-002
February 2001

Data Quality Objectives (DQOs) and Model Development for Relating Federal Reference Method (FRM) and Continuous $PM_{2.5}$ Measurements to Report an Air Quality Index (AQI)

February, 2001

EPA Disclaimer

The information in this document has been funded wholly or in part by the United States Environmental Protection Agency under Contract 68-D-98-030. This document is illustrative guidance which is being distributed as an example of how to relate FRM and continuous PM2.5 measurements to report an Air Quality Index (AQI). The applicable regulations for implementing the AQI can be found in 40 CFR Part 58.50 and Appendix G to Part 58. This document does not substitute for those provisions or regulations, nor is it a regulation itself. Thus, it does not impose binding, enforceable requirements on State or local agencies, and may not apply to a particular situation based upon the circumstances. EPA and State or local decision makers retain the discretion to adopt approaches on a case-by-case basis that differ from this guidance, where appropriate. Therefore, interested parties are free to raise questions and objections about the appropriateness of the application of this guidance to a particular situation; EPA will, and States and local agencies should, consider whether or not the recommendations in the guidance are appropriate in that situation. This guidance is a living document and may be revised periodically without public notice. EPA welcomes public comments on this document at any time and will consider those comments in any future revision of this guidance document.

EPA-454/R-01-002

DATA QUALITY OBJECTIVES (DQOs) AND MODEL DEVELOPMENT FOR RELATING FEDERAL REFERENCE METHOD (FRM) AND CONTINUOUS $PM_{2.5}$ MEASUREMENTS TO REPORT AN AIR QUALITY INDEX (AQI)

Shelly Eberly - U.S. EPA
Terence Fitz-Simons - U.S. EPA
Tim Hanley - U.S. EPA
Lewis Weinstock - Forsyth County NC, Environmental Affairs Department
Tom Tamanini - Hillsborough County FL, Environmental Protection Commission
Ginger Denniston - Texas Natural Resource Conservation Commission
Bryan Lambath - Texas Natural Resource Conservation Commission
Ed Michel - Texas Natural Resource Conservation Commission
Steve Bortnick - Battelle Memorial Institute

U.S. ENVIRONMENTAL PROTECTION AGENCY
Office of Air and Radiation
Office of Air Quality Planning and Standards
Research Triangle Park, North Carolina 27711

February, 2001

Table of Contents

List of Tables

List of Figures

EXECUTIVE SUMMARY

According to Part 58.50 of 40 CFR, all Metropolitan Statistical Areas (MSAs) with a population of 350,000 or greater are required to report daily air quality using the Air Quality Index (AQI) to the general public. AQI is calculated from concentrations of five criteria pollutants: ozone (O_3), particulate matter (PM), carbon monoxide (CO), sulfur dioxide (SO_2), and nitrogen dioxide (NO_2). According to Part 58 of 40 CFR, Appendix G, particulate matter measurements from non-Federal Reference Method (FRM) monitors may be used for the purpose of reporting the AQI if a linear relationship between these measurements and reference or equivalent method measurements can be established by statistical linear regression. This report provides guidance to MSA's for establishing a relationship between FRM and continuous $PM_{2.5}$ measurements.

Chapter 2 of this report details the use of the EPA's Data Quality Objectives (DQOs) process to develop a statistical linear regression model relating FRM and continuous $PM_{2.5}$ measurements. Respectively, Tables 2-2 and 2-3 indicate the quantity of data and quality of model required to confidently use continuous $PM_{2.5}$ data, along with the established model, for the timely reporting of an MSA's AQI. Depending on the level of decision errors tolerable to an individual MSA's decision makers, a minimum of 18 days with both FRM and continuous measurements should be used to develop a model. (In some cases many more days of data are required.) With smaller sample sizes to work with (days < 50), an MSA's model should possess an R^2 value (strength of model) of at least 0.78, while larger sample sizes can lead to a required R^2 value as low as 0.71.

Chapter 3 of this report offers step-by-step guidance to MSA's for developing a regression model relating FRM and continuous $PM_{2.5}$ measurements. Provided is a discussion of data issues likely to be encountered and methods to address them. Real-world examples are used for illustration, and are based on data from Davenport-Moline-Rock Island, IA-IL; Greensboro–Winston-Salem–High Point, NC; Salt Lake City-Ogden, UT; and Houston, TX.

1.0 INTRODUCTION

According to Part 58.50 of 40 CFR, all Metropolitan Statistical Areas (MSAs) with a population of 350,000 or greater are required to report daily air quality using the Air Quality Index (AQI) to the general public. The AQI is calculated from concentrations of five criteria pollutants: ozone (O_3), particulate matter (PM), carbon monoxide (CO), sulfur dioxide (SO_2), and nitrogen dioxide (NO_2). The concentration data used in the calculation are from the State and Local Air Monitoring Stations (SLAMS) required under Part 58 of 40 CFR for each pollutant except PM.

According to Part 58 of 40 CFR, Appendix G, particle measurements from non-Federal Reference Method (FRM) monitors may be used for the purpose of reporting the AQI if a linear relationship between these measurements and reference or equivalent method measurements can be established by statistical linear regression. In fact, some areas already use non-Federal Reference Method (FRM) monitors for the purpose of reporting the AQI and EPA encourages the use of continuous measurements for the sake of timely reporting of the AQI. We recognize, however, that it might not be feasible to find a satisfactory correlation between continuous measurements and FRM measurements of $PM_{2.5}$ in some areas or under some conditions. Air pollution control authorities should not use continuous methods for reporting the AQI in these circumstances.

This document describes the use of continuous $PM_{2.5}$ measurements for the purpose of reporting the AQI, through the establishment of a linear relationship between FRM and continuous $PM_{2.5}$ measurements using statistical linear regression. The document also describes using statistical linear regression to transform continuous PM2.5 measurements into FRM-like data. While not a regulatory requirement, such data transformations might be necessary to report the AQI accurately. There are approximately 240 sites in the $PM_{2.5}$ continuous network, with most of the monitors in the large MSAs. To determine an appropriate model of the relationship between FRM and continuous $PM_{2.5}$ measurements, EPA makes use of the Data Quality Objectives (DQO) process, a seven-step strategic planning approach based on the Scientific Method. The seven-step DQO process is summarized as follows:

1. State the problem.
2. Identify the decision.
3. Identify inputs to the decision.
4. Define the study boundaries.
5. Develop a decision rule.
6. Specify limits on decision errors.
7. Optimize the design for obtaining data.

In general, the DQO process represents a scientific approach to determining the most appropriate data type, quality, quantity and synthesis (i.e., model development) for a given activity (i.e., non-FRM AQI reporting).

This document summarizes the DQO process that was conducted for developing acceptable models to report an AQI using non-FRM continuous $PM_{2.5}$ monitoring data (Chapter 2). Also provided is a "handbook" to guide MSAs in developing their own specific models (Chapter 3). Issues associated with model development are highlighted through four case studies detailed in Appendix B. In particular, data from Davenport-Moline-Rock Island, IA-IL; Greensboro–Winston-Salem–High Point, NC; Salt Lake City-Ogden, UT; and Houston, TX were used as case studies to (1) conduct the DQO process, (2) demonstrate the need for MSA-specific model development, and (3) provide examples of approaches to model development. Table 1-1 summarizes the FRM and continuous $PM_{2.5}$ monitoring data used in this effort.

Table 1-1. Data available for continuous $PM_{2.5}$ DQO development (as of 06/19/00)

MSA	State	Site	FRM PM2.5 (ug/cu meter (Local Conditions))				Continuous PM2.5 (ug/cu meter (Local Conditions))			
			Method	Frequency	Period	n	Method	Frequency	Period	n
Davenport-Moline-Rock Island, Iowa-Illinois	IA	191630015	R Gravimetric	1 in 3 days for 1999 and daily for 2000	01/99-04/00	231	Automated TEOM Gravimetric	Hourly	02/99-04/00	442
		191630013					Automated TEOM Gravimetric	Hourly	01/99-04/00	465
		191630017					Automated TEOM Gravimetric	Hourly	01/99-04/00	478
		191630018	R Gravimetric	1 in 3 days	07/99-04/00	102				
	IL	171610003	Anderson Gravimetric in 1999 and R Gravimetric in 2000	1 in 6 days	01/99-03/00	72				
Greensboro-Winston-Salem-High Point, North Carolina	NC	370670022	R Gravimetric	daily	01/99-03/00	409	Automated TEOM Gravimetric	Hourly	06/99-02/00	259
		370010002	R Gravimetric	1 in 3 days	01/99-09/99	76				
		370570002	R Gravimetric	1 in 3 days	01/99-09/99	78				
		370670024	R Gravimetric	1 in 3 days	01/99-03/00	137				
		370810009	R Gravimetric	daily	01/99-09/99	220				
		370811005	R Gravimetric	1 in 3 days	01, 03, 04, 06-09/00	52				

Note: Continuous PM2.5 measurements were converted from HOURLY to DAILY by taking the average of measurements collected from 1am to midnight.

Table 1-1. Data available for continuous $PM_{2.5}$ DQO development (as of 06/19/00) (continued)

MSA	State	Site	FRM PM2.5 (ug/cu meter (Local Conditions))				Continuous PM2.5 (ug/cu meter (Local Conditions))			
			Method	Frequency	Period	n	Method	Frequency	Period	n
Salt Lake City, Utah	UT	490110001	R Gravimetric	1 in 3 days	01/99 - 03/00	147	TEOM	hourly	12/99 - 07/00	235
		490350003	R Gravimetric	1 in 3 days	01/99 - 03/00	146				
		490350012	R Gravimetric	1 in 3 days	01/99 - 03/00	144				
		490353006	T Gravimetric in 1999 and Met Gravimetric and Anderson Gravimetric in 2000	every day	01/99 - 03/00	403	TEOM	hourly	12/99 - 07/00	212
		490353007	R Gravimetric	1 in 3 days	01/99 - 03/00	133				
		490450002	R Gravimetric	1 in 3 days	01/99 - 03/00	130				
		490494001	R Gravimetric	every day	01/99 - 03/00	417				
		490495010	R Gravimetric	1 in 3 days	01/99 - 03/00	140				
		490570001	R Gravimetric	1 in 3 days	01/99 - 03/00	130				
		490570007	R Gravimetric	1 in 3 days	01/99 - 03/00	130				
Houston, Texas	TX	482011035	R Gravimetric	every day	02/00 - 06/00	109				
		482010026	R Gravimetric	every day	02/00 - 06/00	86	TEOM	hourly	02/00 - 06/00	147
		482010062	R Gravimetric	1 in 3 days	02/00 - 06/00	43				
		482010051	R Gravimetric	1 in 3 days	02/00 - 06/00	41				
		482011039	R Gravimetric	1 in 3 days	02/00 - 06/00	40	TEOM	hourly	03/00 - 06/00	118
		482011037	R Gravimetric	1 in 3 days	02/00 - 06/00	38				
		383390089	R Gravimetric	1 in 3 days	02/00 - 06/00	31	TEOM	hourly	02/00 - 06/00	134
		482011034					TEOM	hourly	02/00 - 06/00	147

Notes:
1. Continuous PM2.5 measurements were converted from HOURLY to DAILY by taking the average of measurements collected from 1am to midnight.
2. Utah sites 490050004 and 490495008 contained only 14 and 4 FRM observations, respectively.
3. Utah sites 490450002, 490494001, 490495008, and 490495010 are not located within the Salt Lake City MSA.

2.0 DATA QUALITY OBJECTIVE (DQO) PROCESS FOR MODEL DEVELOPMENT TO REPORT AN AIR QUALITY INDEX (AQI) WITH CONTINUOUS $PM_{2.5}$ MONITORING DATA

This chapter details the DQO process for establishing a relationship between Federal Reference Method (FRM) $PM_{2.5}$ and continuous $PM_{2.5}$ monitoring data. Each of the seven sections of this chapter corresponds to one of the seven steps of the DQO process. These sections describe the activities conducted and decisions made under each step. The approach is consistent with the EPA Quality Staff report, "Guidance for the Data Quality Objectives Process," EPA QA/G-4, September 1994. Note that the DQO process is recommended by EPA as a tool for model development. The purpose of using this process is to minimize the likelihood of making errors during model development, and ultimately to correctly decide whether the model is adequate for its intended use.

2.1 STEP 1 - STATE THE PROBLEM

The purpose of this step is to define the problem at hand. Activities and outputs from this step include (1) listing planning team members and identifying the decision maker, (2) developing a concise description of the problem, and (3) summarizing available resources and relevant deadlines for the study.

Table 2-1 summarizes the planning team members who participated in this DQO exercise. Communication among planning team members was facilitated mainly through regular conference calls. A concise description of the problem is as follows:

Table 2-1. FRM versus continuous $PM_{2.5}$ model development DQO planning team

Name	Address	Phone Number	Electronic Mail
Decision Makers			
Ginger Denniston	TNRCC P.O. Box 13087 Austin, TX 78711-3087	(512) 239-1673	gdennist@tnrcc.state.tx.us
Terence Fitz-Simons	USEPA/OAQPS AQTAG (MD-14) Research Triangle Park, NC 27711	(919) 541-0889	Fitz-Simons.Terence@epamail.epa.gov
Tim Hanley	USEPA/OAQPS MQAG (MD-14) Research Triangle Park, NC 27711	(919) 541-4417	Hanley.Tim@epamail.epa.gov
Bryan Lambeth	TNRCC P.O. Box 13087 Austin, TX 78711-3087	(512) 239-1657	blambeth@tnrcc.state.tx.us
Ed Michel	TNRCC P.O. Box 13087 Austin, TX 78711-3087	(512) 239-1384	emichel@tnrcc.state.tx.us
Lewis Weinstock	Forsyth County Environmental Affairs 537 North Spruce Street Winston-Salem, NC 27101-1362	(336) 727-8060	weinstl1@co.forsyth.nc.us
Tom Tamanini	Environmental Protection Commission 1410 N. 21st Street Tampa, FL 33605	(813) 272-5530	tamanini@epcjanus.epchc.org
Primary Contractor Contact			
Steve Bortnick	Battelle 505 King Avenue Columbus, OH 43201-2693	(614) 424-7487	bortnick@battelle.org
Primary EPA Contact			
Shelly Eberly	USEPA/OAQPS MQAG (MD-14) Research Triangle Park, NC 27711	(919) 541-4128	Eberly.Shelly@epamail.epa.gov

Problem Statement: *It is desired to use continuous $PM_{2.5}$ measurements for the purpose of reporting an Air Quality Index (AQI). According to Part 58 of 40 CFR, Appendix G, these data may be used for this purpose if a linear relationship between continuous measurements and reference or equivalent $PM_{2.5}$ method measurements can be established by statistical linear regression. Therefore, a model relating FRM and continuous $PM_{2.5}$ measurements, possibly adjusting for meteorological data, is required.*

In general, the resources and deadlines for establishing the relationship referred to in the above problem statement will vary from one MSA to another. Resource and time constraints should be specified in the early stages of this process.

2.2 STEP 2 - IDENTIFY THE DECISION

The purpose of this step is to clearly define the decision statement the study will attempt to resolve. Activities include (1) identifying the principal study question, (2) defining the alternative actions that could result from resolution of the principal study question, (3) combining the principal study question and the alternative actions into a decision statement, and, if necessary, (4) organizing multiple decisions. The expected output from this step is a decision statement that links the principal study question to possible actions that will solve the problem.

The principal activity associated with the overall DQO exercise is the development of a model relating FRM $PM_{2.5}$ measurements with continuous $PM_{2.5}$ measurements, so that continuous data can be used for the purpose of reporting an AQI or transformed into FRM-like data for the purpose of reporting an AQI. For the purposes of this document, EPA assumes that transformed data will more accurately estimate FRM data than un-transformed data. The principal issue, therefore, is the determination of whether the model that is ultimately derived is acceptable. If the model is deemed acceptable, an MSA's AQI may be reported on a

more timely basis using continuous $PM_{2.5}$ data. If not, given the potential consequences (see DQO Step 6), the model should not be used, which leads to the conclusion (possibly temporary) that the MSA's AQI should not be reported using continuous $PM_{2.5}$ data. Further investigation might be conducted to obtain an acceptable model, such as developing alternative models, evaluating the continuous and/or FRM monitoring methods (e.g., revisit the associated Quality Assessment Project Plan), or waiting for more data to re-apply the current model. This leads to the following:

Decision Statement: *Is the statistical linear model relating FRM $PM_{2.5}$ measurements to continuous $PM_{2.5}$ measurements acceptable for transforming continuous measurements for the purpose of reporting the MSA's AQI? If yes, then the continuous $PM_{2.5}$ data, along with the model, can be used to report the MSA's AQI. If no, do not use continuous $PM_{2.5}$ data to report the MSA's AQI. In the latter case, an MSA might attempt to improve the model until it is acceptable. If this fails, evaluation of the continuous and/or FRM monitoring methods may be necessary.*

2.3 STEP 3 - IDENTIFY INPUTS

The purpose of this step is to identify the informational inputs needed to resolve the decision statement and determine the inputs that require environmental measurements. Activities include (1) identifying the information required to resolve the decision statement, (2) determining the sources for each item of information identified, (3) identifying the information necessary to establish the action level, and (4) confirming that appropriate analytical methods exist to provide the necessary data. The expected outputs from this step are the list of informational inputs needed for the resolution of the decision statement and the list of environmental variables or characteristics to be measured in the study.

The list of environmental measurements required for this study are as follows:

- FRM $PM_{2.5}$ daily measurements,
- continuous $PM_{2.5}$ hourly measurements, and possibly
- meteorological data such as temperature.

At the most basic level, the MSA will require a set of days for which both FRM $PM_{2.5}$ measurements and continuous $PM_{2.5}$ measurements have been obtained from sites within the MSA. Such information is obviously vital to developing a model relating the two measures. Ideally, (1) a large number of days will be available, including data spanning at least one year, (2) at least some of the FRM-continuous data will be co-located, and (3) meteorological data will be available for model improvement. In many cases, these data will be available in AIRS. In some cases, data will be accessible from an MSA's archive in spreadsheet or other format. Along with data, guidelines for the approach to model development are available from most introductory statistical linear regression texts. Guidance specifically tailored to the problem at hand is provided in Chapter 3 of this report.

For this problem, there is no regulatory threshold value around which a decision-making action level might be defined. Therefore, the expert opinion of veteran data analysts will be solicited to determine a measure and associated action level around which model adequacy can be determined.

2.4 STEP 4 - DEFINE THE STUDY BOUNDARIES

The purpose of this step is to define the spatial and temporal boundaries covered by the decision statement. Activities include (1) specifying the characteristics that define the population of interest, (2) defining the geographical area within which all decisions must apply, (3) when appropriate, dividing the population into strata that have relatively homogeneous characteristics, (4) determining the time frame to which the decision applies, (5) determining when to collect data, (6) defining the scale of decision making, and (7) identifying any practical constraints on data collection. The expected outputs from this step are a detailed description of the spatial and temporal boundaries of the problem along with a summary of the practical constraints that may interfere with the study.

The population of interest is daily $PM_{2.5}$ concentrations for the MSA, measured in $\mu g/m^3$. The MSA is the geographical area within which the decision that the model is or is not acceptable is to be applied. The time frame to which the decision applies will be up to individual MSA decision makers. The recommendation is that an acceptable model should be checked for accuracy and updated if necessary at least yearly or, better yet, quarterly. Hence, the time frame to which the decision applies is, starting at the time of model acceptance, the upcoming 90-day to one year period.

Data permitting, some MSAs might develop models specific to sub-regions within the MSA; hence the spatial scale of decision making could be anywhere from an MSA sub-region surrounding the site(s) used to develop the model up to the entire MSA itself. The temporal scale of decision making might range from a few days (if a model is updated or replaced) up to an entire year (if the MSA decision makers feel the model is still accurate a year after development).

It is assumed that both FRM and continuous data are already being collected according to a regular sampling schedule. Therefore, in most cases, the MSA's current and historical monitoring and sampling infrastructure will impose the most significant practical constraint on data collection. The MSA might decide to modify sampling, if resources permit, to improve its ability to build the relation between FRM and continuous $PM_{2.5}$ monitoring data.

2.5 STEP 5 - DEVELOP A DECISION RULE

The purpose of this step is to define the parameter of interest, specify the action level, and integrate previous DQO outputs into a single statement that describes a logical basis for choosing among alternative actions. Activities and expected outputs include (1) specifying the statistical parameter that characterizes the population, (2) specifying the action level for the study, and (3) combining the outputs of the previous DQO steps into an "if...then..." decision rule that defines the conditions that would cause the decision maker to choose among alternatives.

Since the purpose of this exercise is to develop an acceptable model that relates FRM and continuous $PM_{2.5}$ measurements, DQO planning team members determined that the statistical parameter of interest is the R^2 parameter provided as standard output from all software packages that perform statistical linear regression. In general, R^2 measures the strength of the model fit to the data. In this case, R^2 measures the correlation between measured and modeled FRM $PM_{2.5}$ data.

In simple regression (i.e., regression of FRM on continuous $PM_{2.5}$ data with no adjustment for seasonality, MET data, etc.), R^2 is simply the square of the correlation coefficient between FRM and continuous $PM_{2.5}$ measurements. In multiple regression (i.e., regression of FRM on continuous $PM_{2.5}$ data along with other variables such as seasonality, MET data, etc.), R^2 is known as the multiple correlation coefficient or coefficient of multiple determination, and its interpretation is less straightforward. In either case, simple or multiple regression, R^2 is the square of the correlation coefficient between observed FRM $PM_{2.5}$ data values and their modeled counterparts, as derived from a fitted statistical linear model using continuous data. This latter interpretation is the basis for establishing DQOs for the model to be developed and the data used in that development.

Suppose there are *n* days of FRM and continuous $PM_{2.5}$ data for use in model development. Define y_i to be the FRM concentration on the ith day, $\hat{y}_i$ to be the modeled FRM concentration on the ith day, and $\bar{y}$ to be the average of the *n* FRM measurements. Then the formula for R^2 can be written as follows:

$$R^2 = \frac{\sum_{i=1}^{n} \left(\hat{y}_i - \bar{y}\right)^2}{\sum_{i=1}^{n} \left(y_i - \bar{y}\right)^2}$$

which indicates that R^2 measures the proportion of total variation in FRM data explained by the model (i.e., how well the model fits the data).

The action level around which a model might be deemed acceptable was determined by DQO planning team members to be the value of R^2 equal to 0.60. At first, this action level might appear somewhat lax to data analysts used to interpreting strong regression relationships as those with an R^2 value in the range of 0.80 or above. However, it is important to keep in mind that in the current context a decision is to be made based on estimating the model's true R^2 value, a rather uncommon activity in practice. In most applied contexts, the sample statistic R^2 obtained from software regression output is treated as the true R^2 value, when in fact it is only an estimate of the true unknown value. Under a hypothesis testing scenario, accepting or rejecting a model based on a true R^2 action level of 0.60 is shown in Table 2-3 of Section 2.7 as equivalent to requiring a sample R^2 value equal to around 0.80, a model adequacy threshold more common to most applied data analysts. Furthermore, a true R^2 value of 0.60 is equivalent to a true correlation coefficient of 0.77 between observed and modeled FRM $PM_{2.5}$ data, which is a rather strong correlation that indicates good agreement between actual data and model predictions (i.e., a good model fit).

The above discussion leads to the following:

"If...then..." Statement: *If the true R^2 value from the statistical linear regression model relating FRM and continuous $PM_{2.5}$ measurements within the MSA over the next 90-day to one year period is greater than 0.60, then continuous $PM_{2.5}$ data can be used, along with the model, to report the MSA's AQI. Otherwise, the model in its current form is not acceptable, so continuous $PM_{2.5}$ data should not be used for this purpose.*

2.6 STEP 6 - SPECIFY TOLERABLE LIMITS ON DECISION ERRORS

The purpose of this step is to specify the decision maker's tolerable limits on decision errors. Activities include (1) determining the possible range of the parameter of interest, (2) identifying the decision errors and choosing the null hypothesis, (3) specifying the range of possible parameter values where the consequences of decision errors are relatively minor (in the gray region), and (4) assigning the probability values to points above and below the action level that reflect the tolerable probability for the

decision maker's tolerable decision error rates based on a consideration of the consequences of making an incorrect decision.

As stated in DQO Step 5 above, the correlation between observed and modeled FRM $PM_{2.5}$ values (or R^2) is a measure of the model's adequacy, and DQO planning team members determined that a model is acceptable if its true R^2 value is at or above the action level of 0.60. Hence, the decision as to whether the model is acceptable is statistically formalized as the following hypothesis test:

$$H_0: R^2 \# 0.60 \text{ versus } H_a: R^2 > 0.60 ;$$

where, overall, R^2 values can theoretically range from 0.0 (i.e., no relation between actual and modeled FRM $PM_{2.5}$ measurements) to 1.0 (i.e., perfect correlation between actual and modeled FRM $PM_{2.5}$ measurements).

The null or baseline hypothesis of $R^2 \# 0.60$ is chosen because the decision error associated with this conclusion is considered to be the most serious, and thus should be guarded against. Specifically, a false rejection decision error that the model is adequate ($R^2 > 0.60$) when in fact it is not ($R^2 \# 0.60$) could result in misleading AQI reporting in the form of incorrectly claiming either good or bad air quality. In contrast, the false acceptance decision error that the model is unsatisfactory ($R^2 \# 0.60$) when in fact it is adequate ($R^2 > 0.60$) simply results in not using (or delaying the use of) continuous $PM_{2.5}$ measurements and the associated model to report the AQI.

Along with the above hypothesis statement, three additional parameters must be specified in order to formally accept or reject the model; namely, the false rejection decision error rate (", the false acceptance decision error rate ($), and the size of the gray region in decision making ()). The false rejection decision error rate (") specifies the maximum probability of claiming the model is adequate ($R^2>0.60$) when in fact it is not. Common values for " are 0.0l, 0.05, 0.10, and 0.20. The chosen level of " will depend on the degree to which individual MSA decision makers wish to protect against false rejection decision errors. Smaller " values are more restrictive and demand a better model along with more data for establishing that model.

The false acceptance decision error rate ($) specifies the maximum probability of claiming the model is not adequate ($R^2 \leq 0.60$) when in fact it is ($R^2 > 0.60$). Common values for $ are 0.20, 0.30, and 0.40. The chosen level of $ will depend on the degree to which individual MSA decision makers wish to protect against false acceptance decision errors. Smaller $ values are more restrictive and demand a better model along with more data for establishing that model.

The size of the gray region in decision making (() specifies an area, starting at $R^2 = 0.60$ up to $R^2 = (0.60 + ()$, within which somewhat higher false acceptance decision error rates ($) are considered tolerable. Allowing for a gray region in decision making is necessary given that real-world data are imperfect, and, therefore, do not lead to extremely confident decision making very near an action level of concern (in this case, just above $R^2 = 0.60$). There are no common values for (, as its specification will depend on the problem at hand. In this case, given that the action level is set at $R^2 = 0.60$, (values in the range of 0.20, 0.25, and 0.30 would appear appropriate. These (values lead to gray regions of (0.60-0.80), (0.60-0.85), and (0.60-0.90), respectively. As with $, the chosen level of (will depend on the degree to which individual MSA decision makers wish to protect against false acceptance decision errors. Smaller (values are more restrictive and demand a better model along with more data for establishing that model.

As an example, consider the DQO parameters " = 0.05, $ = 0.30, and (= 0.25. Figure 2-1 provides a visual interpretation of the meaning of each of these parameters. The figure draws a curve indicating the probability of claiming the true R^2 value is above the action level of 0.60 (vertical axis) as a function of the true unknown R^2 value (horizontal axis). Notice that for all values of $R^2 \leq 0.60$, the curve remains below the 0.05 threshold on the vertical axis. In other words, if the model is truly inadequate ($R^2 \leq 0.60$), then the chance of claiming otherwise is never more than five percent (i.e., " = 0.05). Likewise, if the model is quite good ($R^2 \geq 0.85$), then the chance of claiming otherwise is never more than thirty percent (i.e., $ = 0.30). Finally, if the model is good, but only marginally so ($0.60 < R^2 \leq 0.85$), then the chance of claiming otherwise could be substantial (i.e., more than 30 percent). Such is the burden of decision making based on imperfect real-world data.

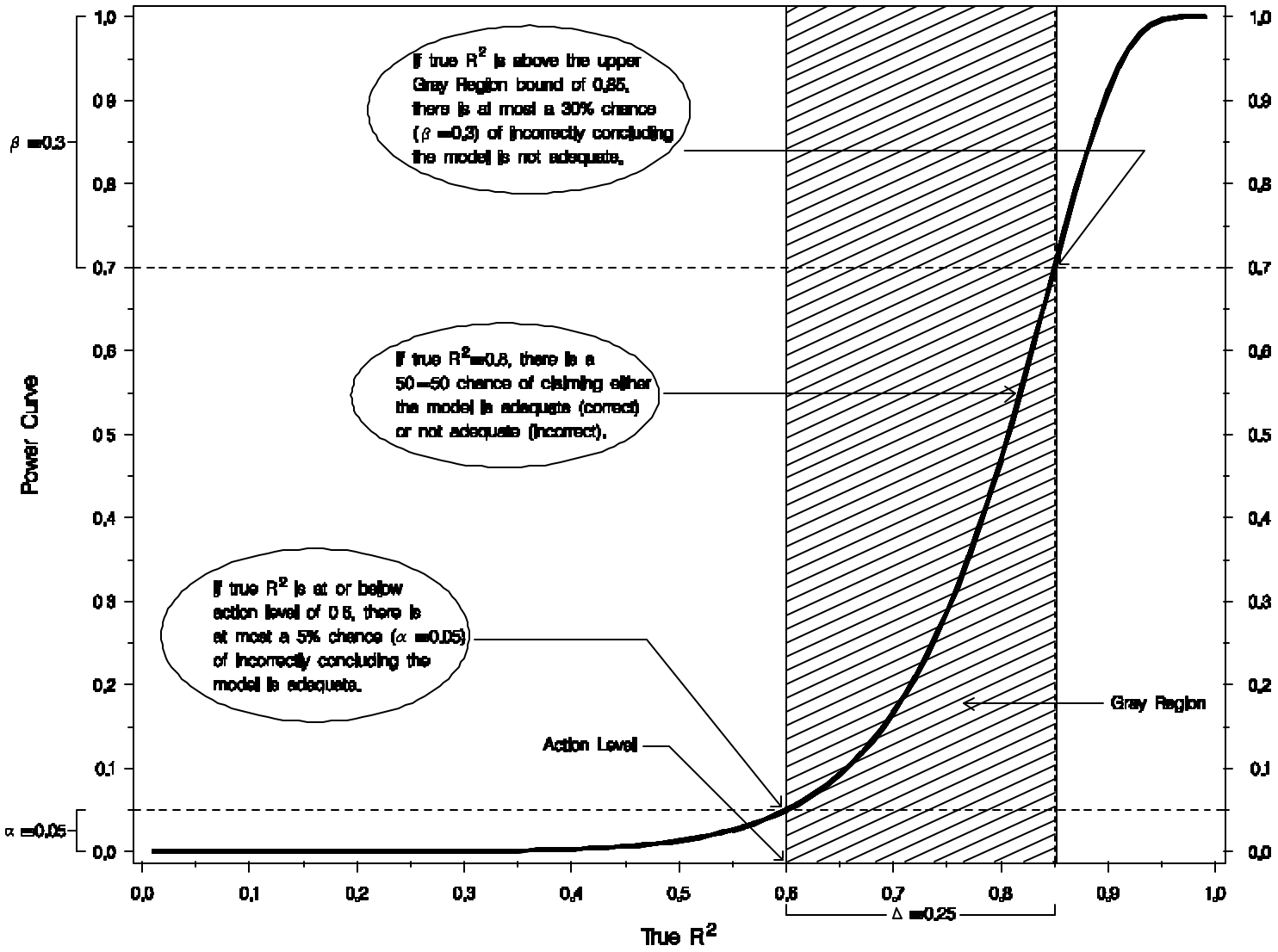

Figure 2-1. Example Decision Curve when N=213, "=0.05, $=0.3, and)=0.25

2.7 STEP 7 - OPTIMIZE THE DESIGN FOR OBTAINING DATA

The purpose of this step is to identify a resource-effective data collection design for generating data that are expected to satisfy the DQOs. Activities include (1) reviewing the DQO outputs and existing environmental data, (2) developing general data collection design alternatives, (3) formulating the mathematical expressions needed to solve the design problems for each design alternative, (4) selecting the optimal sample size that satisfies the DQOs for each design alternative, (5) selecting the most resource-effective design that satisfies all of the DQOs, and (6) documenting the operational details and theoretical assumptions of the selected design in the sampling and analysis plan. The expected output from this step is the most resource- effective design for the study that is expected to achieve the DQOs.

The purpose of this DQO exercise was to provide guidelines for MSAs that would like to use continuous $PM_{2.5}$ monitors for timely reporting of their AQI. The purpose was not to determine the exact model or type and amount of data to be used by each MSA. As such, Step 7 of the DQO process in this case is intended to provide a range of data scenarios and DQO parameter specifications under which MSAs might develop a model relating FRM and continuous $PM_{2.5}$ measurements. Chapter 3 of this report provides further detail on the approach to model development and important issues that must be considered.

Using the parameter of interest and action level defined in Step 5 along with the range of reasonable decision errors and gray regions defined in Step 6, Table 2-2 presents a range of sample size requirements sufficient to confirm a model as adequate or otherwise. Table 2-3 presents a lower bound on the associated sample R^2 value (i.e., the R^2 value that is output from software used to fit the model) that is required in order to decide the model is adequate. The shaded cells of Tables 2-2 and 2-3 correspond to sample sizes either too small to be recommended ($n<18$) or too large to be practical ($n>730$, or two full years of daily data). Appendix A provides the statistical details and assumptions used in deriving these two tables.

Table 2-2. Sample size requirements for model development by ", $, and) under a null hypothesis of H_0: R^2 # 0.6

Size of Gray Region ())	False Acceptance Decision Error ($)	False Rejection Decision Error (")			
		0.20	0.10	0.05	0.01
0.30	0.40		18	37	119
	0.30		31	61	176
	0.20	24	57	103	266
0.25	0.40	18	56	126	422
	0.30	36	104	213	628
	0.20	79	196	365	
0.20	0.40	53	196	454	
	0.30	124	373		
	0.20	282	709		

Table 2-3. Lower bound on observed model R^2 value necessary for concluding model adequacy by ", $, and) under a null hypothesis of H_0: R^2 # 0.6

Size of Gray Region ())	False Acceptance Decision Error ($)	False Rejection Decision Error (")			
		0.20	0.10	0.05	0.01
0.30	0.40		0.87	0.88	0.88
	0.30		0.85	0.86	0.87
	0.20	0.79	0.82	0.84	0.85
0.25	0.40	0.81	0.82	0.83	0.83
	0.30	0.78	0.80	0.81	0.82
	0.20	0.75	0.77	0.79	
0.20	0.40	0.76	0.77	0.78	
	0.30	0.74	0.75		
	0.20	0.71	0.73		

For example, suppose an MSA has around 200^{+} days' worth of co-located FRM and continuous $PM_{2.5}$ measurements, and possibly meteorological data as well, from which to develop a model. Among other choices, Table 2-2 indicates that a test of whether the model is adequate could be done at " = 0.05, $ = 0.30, and) = 0.25. Table 2-3 indicates that under these parameters and with this sample size, the final model would need to achieve an observed R^2 value of 0.81 or higher in order to confidently conclude it is good enough for its intended use. The interpretation of this scenario (200+ observations and R^2 $ 0.81 for model acceptance) is as follows:

- If the model is not good (true R^2 # 0.60), then there is only a 5 percent chance (" = 0.05) of incorrectly concluding the model is good, and hence using it for reporting the AQI.
- If the model is quite good (true $R^2 > 0.85$), then there is only a 30 percent chance of incorrectly concluding the model is not good, and hence not using it for reporting the AQI.
- If the model is only marginally good ($0.60 <$ true R^2 # 0.85), then there is a greater than 30 percent chance of incorrectly concluding the model is not good.

The DQO planning team that developed these guidelines recognizes (as specified in Step 4) that most MSAs will be faced with developing a model based on data already collected. Therefore, as is often the case, many MSAs may choose to use this DQO process in what amounts to its reverse order. Instead of using the process to determine how much data are required, the amount of data an MSA is constrained to can be compared to Table 2-2 to determine exactly what levels of confidence in decision making are obtainable. Based on the MSA's available data and the achievable/chosen cell within Table 2-2, Table 2-3 then provides an answer for the model's R^2 value that must be reached in order to conclude the model is good.

For example, if an MSA has less than 30 days of observations to work with, then Table 2-2 provides three options (i.e., three specifications of ", $, and) corresponding to n = 18 or 24). The MSA can choose from among these three options, then use Table 2-3 to identify the associated R^2 value their model must achieve if it is to be used along with continuous $PM_{2.5}$ data for reporting its AQI.

In conclusion, the DQO planning team that developed these guidelines recommends using Tables 2-2 and 2-3 as an indication of how much data are required in model development and how good the resulting model must be. As Table 2-2 suggests, any MSA that does not possess at least 18 observations for model development probably should not consider the activity until more data become available. Furthermore, although n = 18 observations is displayed in Table 2-2, MSAs with just that amount of data still might conclude that the decision errors associated with such a small sample size are simply too large to warrant conducting the activity at the present time. Finally, few MSAs if any will possess a data set for model development with a sample size that exactly matches Table 2-2. In such cases, reasonable judgment should be used in identifying the cell(s) of Table 2-2 that most closely match the data at hand.

3.0 GUIDELINES FOR MODEL DEVELOPMENT

This chapter contains a series of nine steps to help you develop a model that converts your continuous $PM_{2.5}$ measurements into values associated with your FRM measurements for reporting the AQI based on your measurements from the continuous monitor. The steps also guide you through evaluating the model in both an absolute sense (how to improve the model until it meets your needs) and evaluating the spatial range of validity for your model. Throughout the steps are examples from actually carrying out this process in several MSAs and the special issues that arose. Specifically, four case studies were conducted using data from Davenport-Moline-Rock Island, IA-IL; Greensboro–Winston-Salem–High Point, NC; Salt Lake City-Ogden, UT; and Houston, TX. We expect that the users of this document will be familiar with the measurement process and reporting of the AQI.

Steps 1 through 4 contain the "exploratory analysis." These will help you get the best possible data set to work with and help you determine how much work it may take to get the results that you want. Steps 5 through 7 develop the initial models and evaluate the spatial variability within your MSA. Step 8 details how you might go about improving the model until it meets your needs. Finally, Step 9 takes care of some loose ends that you will need to consider. We have limited the statistical/data analysis procedures to things that can be done with common spreadsheets, such as MS Excel.

Before we start we need to set the stage. What is your objective? This is an important question that different people will answer differently and hence will modify the steps below to meet their needs. Do you want to predict the daily maximum, the average of your core FRMs, just correlate your continuous monitor to a co-located (or nearby) FRM monitor, or "calibrate" each continuous and FRM pair? We suggest that you start with the latter, because this will help determine the spatial range of the predictions that you can get while developing the model itself. Your needs and resources will guide the process that you use.

3.1 STEP 1– IDENTIFY YOUR SOURCES OF DATA AND THE TIME FRAME FOR THE AVAILABLE DATA

The key to this step is to find as much useful data as possible, and this may mean throwing out some of the available data. Ideally, you want to have a long time series of measurements from all the continuous and FRM monitors within your MSA *from the same days*. To understand the spatial variability we will want to compare how well one pair of monitors relates to each other versus another pair. However, this changes from day to day. What we want to avoid as much as possible is basing the comparison for one pair on a set of days with very little variability to another pair that used mostly days with a lot of variability. This must be balanced with simply having enough data to base a model on. Hence if one of the continuous monitors has only been running a month, for example, then you may not want to include this monitor. You also may end up using only every third day of data from a co-located continuous – FRM pair. The priorities are for the co-located monitors and the core FRMs. If you cannot get a set of at least 18 days with all the monitors running, then you need to keep in mind that some of the comparisons may be a little misleading. You will also want a table of the relative distances between each pair.

3.2 STEP 2 – GRAPHICAL EXPLORATION PART ONE

While rarely reported, unless there is a problem, virtually all statistical analyses start with summary statistics and simple box plots and histograms. Start with a histogram or box plot (your choice) of the concentration data from each monitor. Using Utah data, Figure 3-1 provides an example of histograms for comparing continuous with FRM measurements as well as comparing untransformed with log-transformed measurements. What you are looking for are obvious differences between the continuous monitoring data and the FRM data. Some of the things that we found were:

- Concentrations over 9,000 (AIRS null value codes),
- Continuous data taken immediately after operator intervention,
- Negative or zero concentrations from the continuous monitors (when material is volatilizing faster than it is accumulating), and
- Values between 100 and 400 $\mu g/m^3$ (possibly incorrect).

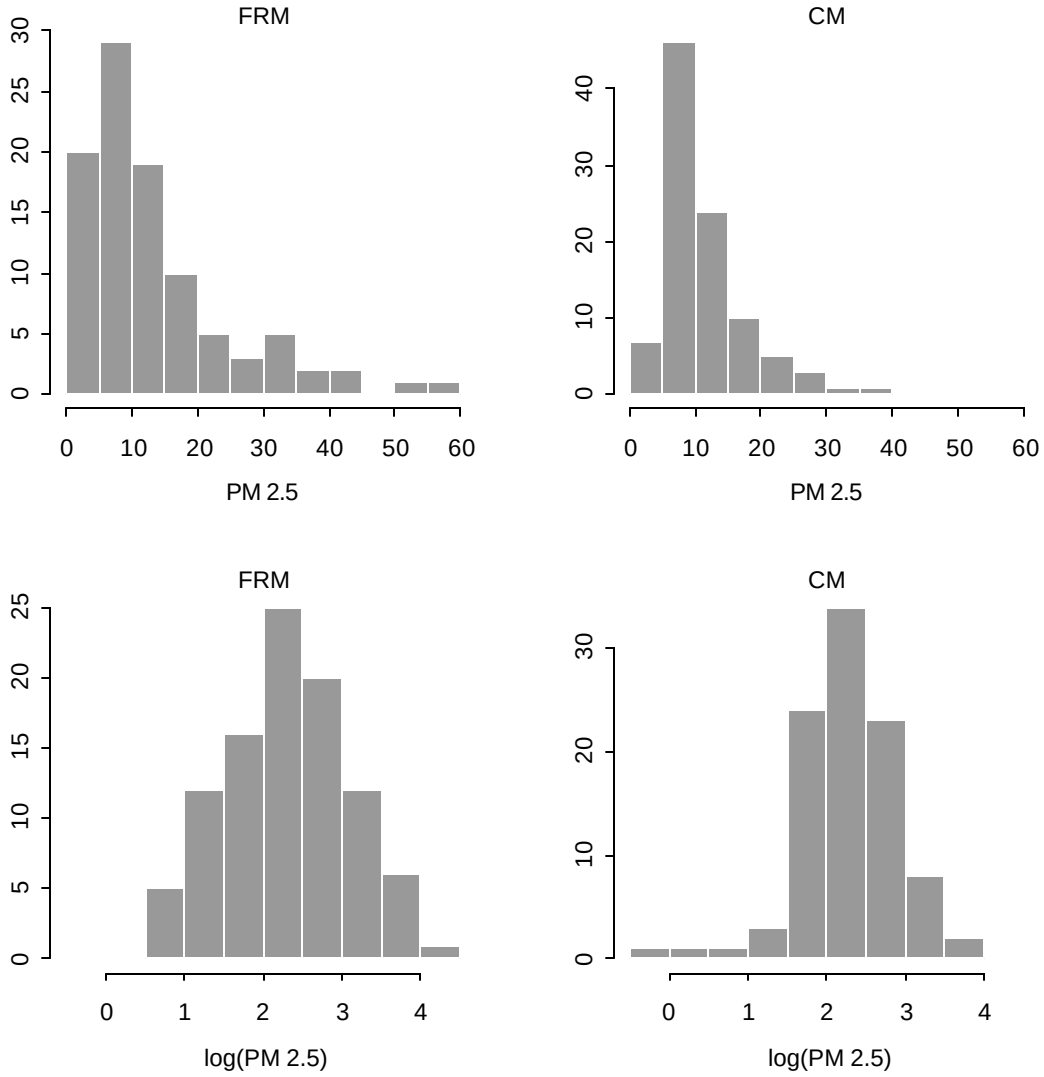

Figure 3-1. Side-by-side histogram summary data from the co-located site 49049001 in Utah. The top two histograms use untransformed data and the bottom two histograms use log-transformed data.

You should also look for unusual patterns and outliers. For example a bimodal pattern could result from a change in the continuous monitor's settings or a big change in the general weather patterns (such as a large precipitation event). You only want data that will be representative of the future values expected for your MSA.

3.3 STEP 3 – PREPARING THE DATA SET

You will need to convert continuous (usually hourly) data into daily averages. We recommend only using days with at least 75 percent completeness. You may also only want days where the FRM data are above (or well above) its MDL, but do not make this requirement so stringent that you lose a significant portion of your data.

Do you need or want to log-transform your data? Go back to your histograms/box plots. Is there a wide range of data with very few high points? One or two (valid) points that are very different from the rest can be very influential in the regression. Suppose you have an isolated point around 50 $:g/m^3$ with a 5 percent measurement error and the rest of the data are around 10 $:g/m^3$. The relatively small errors for the small values will tend to cancel each other, while the single error for the larger value has nothing to average out. The resulting regression line will basically go through the center of the small values and through the larger point. There are two things working against you here, the difference in the absolute size of the errors and a level arm effect. Log-transforming the data can treat both of these problems.

For an example of the benefit of log-transforming your data, consider a site in the Iowa-Illinois MSA with co-located continuous and FRM data. Table 3-1 summarizes the results from the least squares regression, with and without log-transforming the data. Notice the marginal improvement in the R^2 value. More importantly, Figure 3-2 shows how an influential point in the upper right corner is brought closer to the main body of the data when log-transforming (top two plots), how the histogram of the least square residuals become less skewed (middle two plots), and how the spread of the residuals when plotted versus the predicted values becomes more homogenous (bottom two plots). All of these modifications to the data due to log-transforming are improvements toward a more appropriate statistical model.

Table 3-1. Least squares regression summary for Iowa-Illinois MSA co-located continuous and FRM data, untransformed and log-transformed

	N	Intercept	se(Int.)	Slope	se(slope)	R^2	RMSE
untransformed	214	2.661	0.638	0.956	0.050	0.631	4.512
log-transformed	214	0.173	0.106	0.988	0.045	0.693	0.327

Consider a log-transformation if you have isolated large values in your data, or if you suspect that your measurement error is proportional to the size of the response. Our four case studies in Appendix B have been done both ways; only someone familiar with the characteristics of the data can really decide which is most appropriate. If you just do not know, do it both ways and see how much the answers differ.

3.4 STEP 4 – GRAPHICAL EXPLORATION PART TWO

For each continuous-FRM pair that you want to compare, make a scatter plot of the continuous versus FRM values (with the FRM values on the vertical scale). Include a 45-degree line in the plot. A vertical shift from the 45-degree line shows an overall bias. Figure 3-3 demonstrates a consistent bias in the three Texas MSA sites with co-located continuous and FRM data. The solid line is the 45-degree line. If the data tend to cluster around this line, then no overall bias is present. The dashed line is the simple least squares regression line for each associated case. The deviation of the dashed line from the solid line in Figure 3-3 represents the overall bias present in the continuous measurements relative to the FRM measurements.

The degree of scatter of a set of points in a scatter plot shows how correlated the continuous and FRM measurements are with one another. For example, in Figure 3-4, compare the data from the North Carolina MSA to the data from the Iowa-Illinois MSA. Figure 3-4 indicates a much more reliable relationship between continuous and FRM data in the North Carolina MSA relative to the Iowa-Illinois MSA.

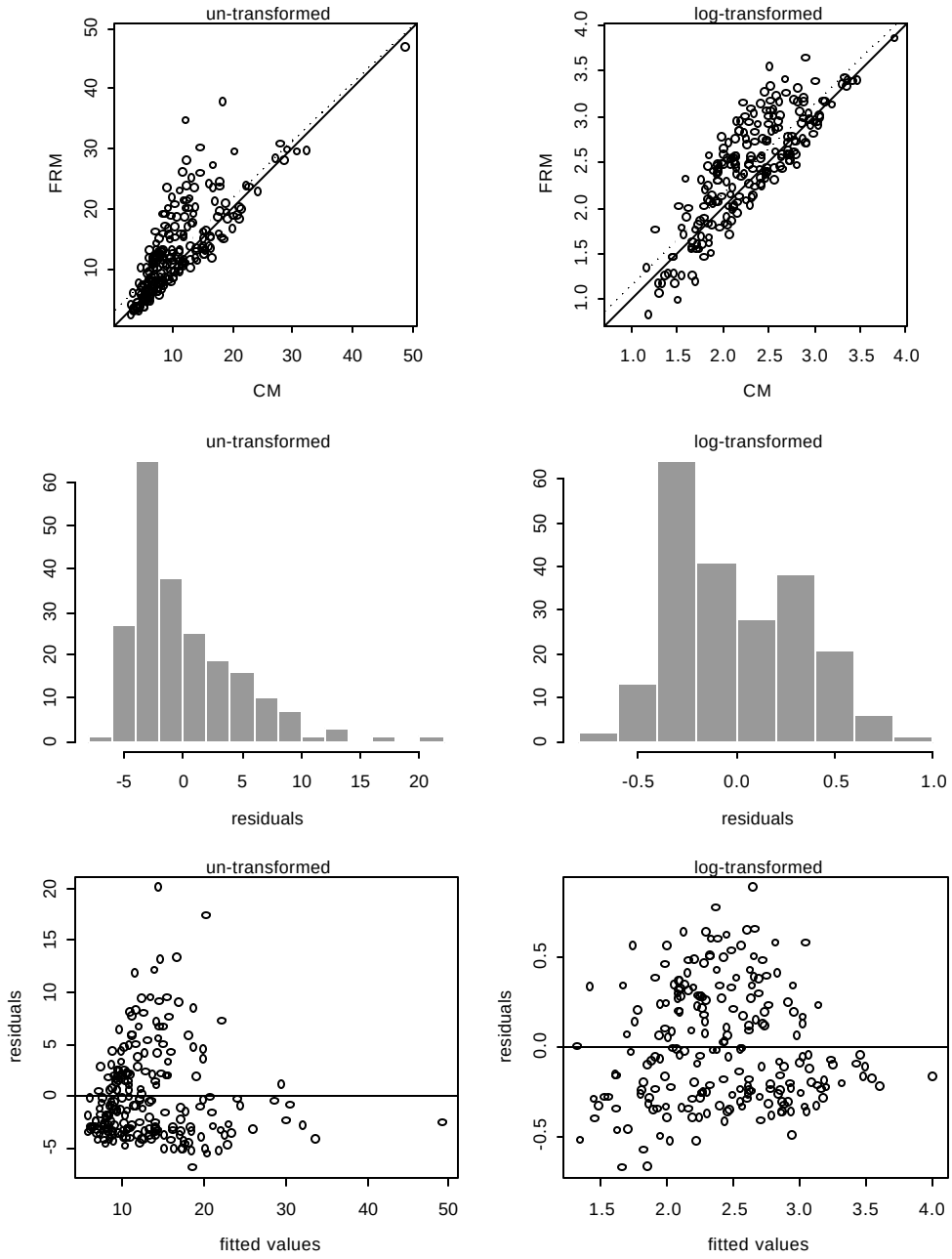

Figure 3-2. **An example of the effect of log-transforming the data. The $PM_{2.5}$ residual concentrations are from a model for the co-located site in the Iowa-Illinois MSA.**

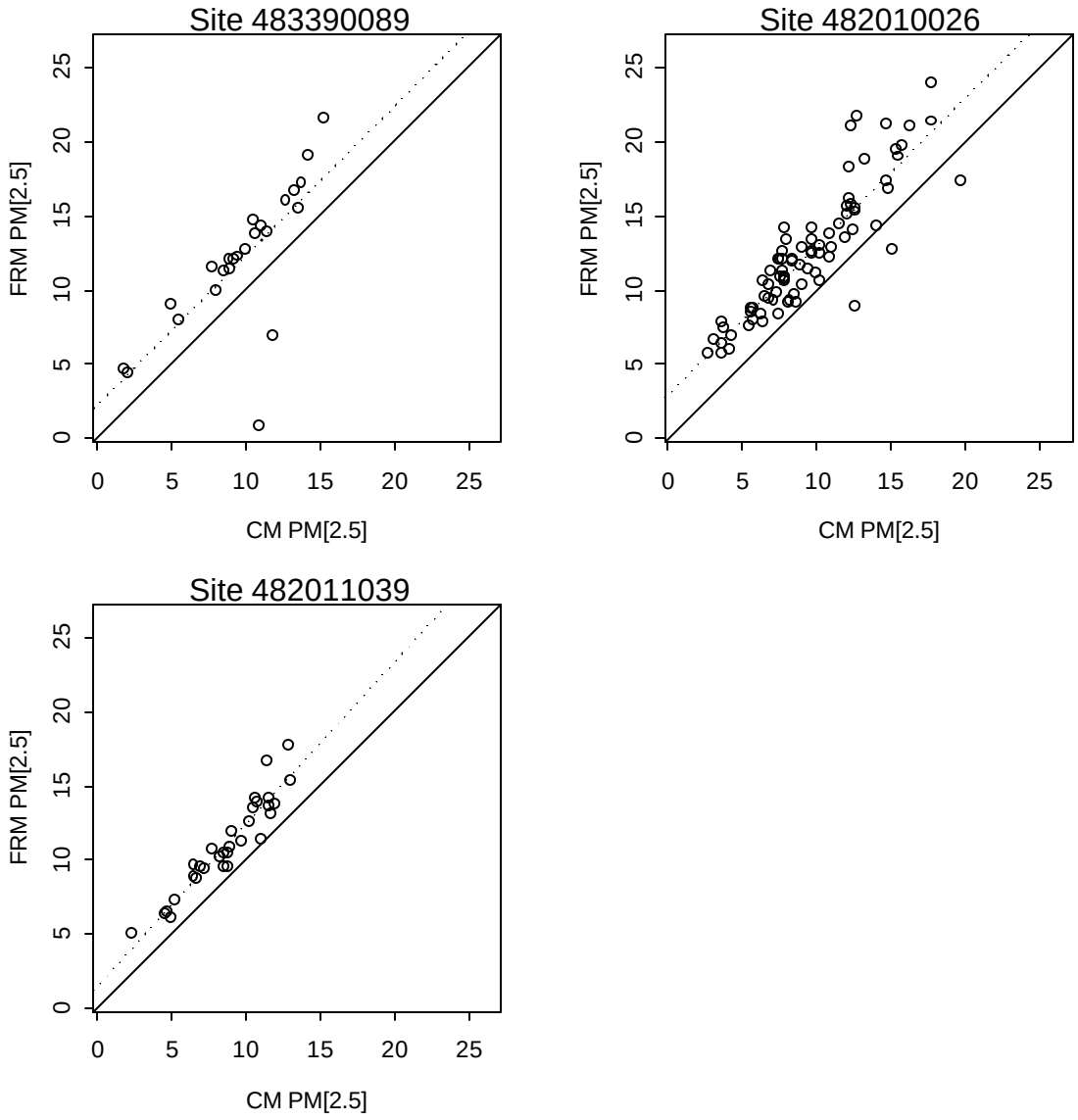

Figure 3-3. **Scatter plot of FRM $PM_{2.5}$ measurements versus continuous $PM_{2.5}$ measurements at the three co-located Texas MSA sites. The solid line shown is the 45 degree line and the dashed line is a regression line.**

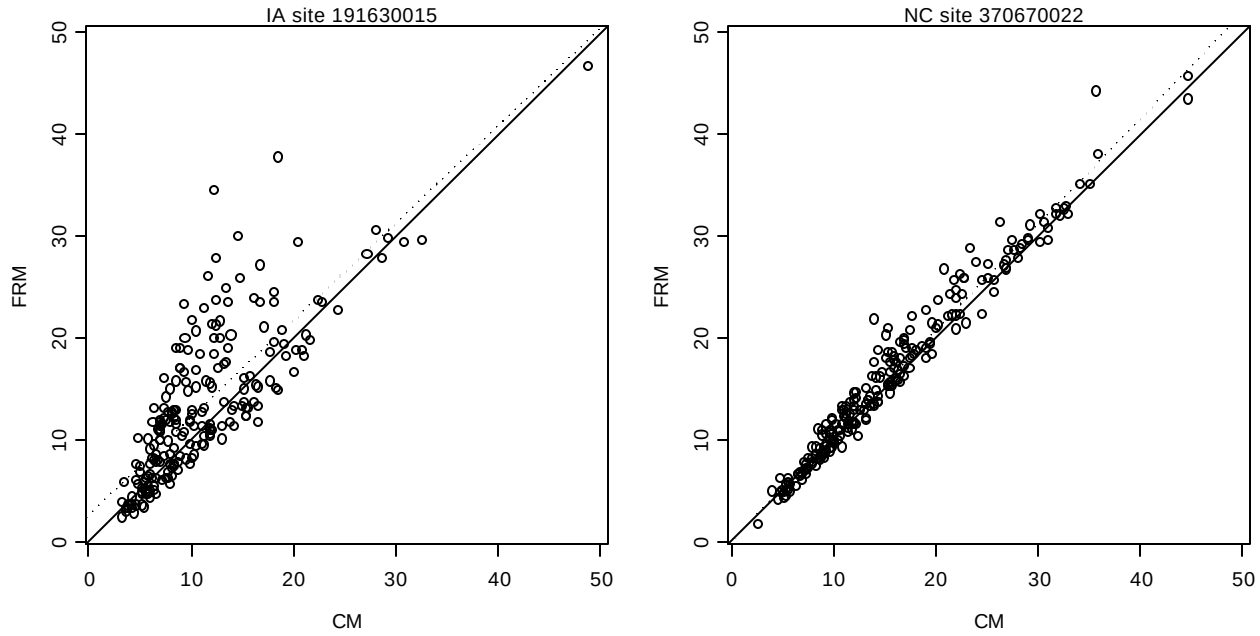

Figure 3-4. An example of different correlations between FRM and continuous measurements; an Iowa-Illinois MSA site to the left and a North Carolina MSA site to the right. The solid line is a 45-degree line and the dashed line is a regression line.

You also want to look for outliers or unusual points, which can strongly influence the regression results. For example Figure 3-5 demonstrates the impact of removing outliers from a regression. The data in Figure 3-5 correspond to a site at the Texas MSA, where the removal of two outliers dramatically improved the R^2 value from 0.54 to 0.95, and marginally impacted the resulting model intercept (2.20 to 1.99) and slope (1.01 to 1.12). Caution should be exercised when removing apparent outliers from a data set. A more careful investigation will often reveal the important circumstances underlying the existence of the outliers in the first place.

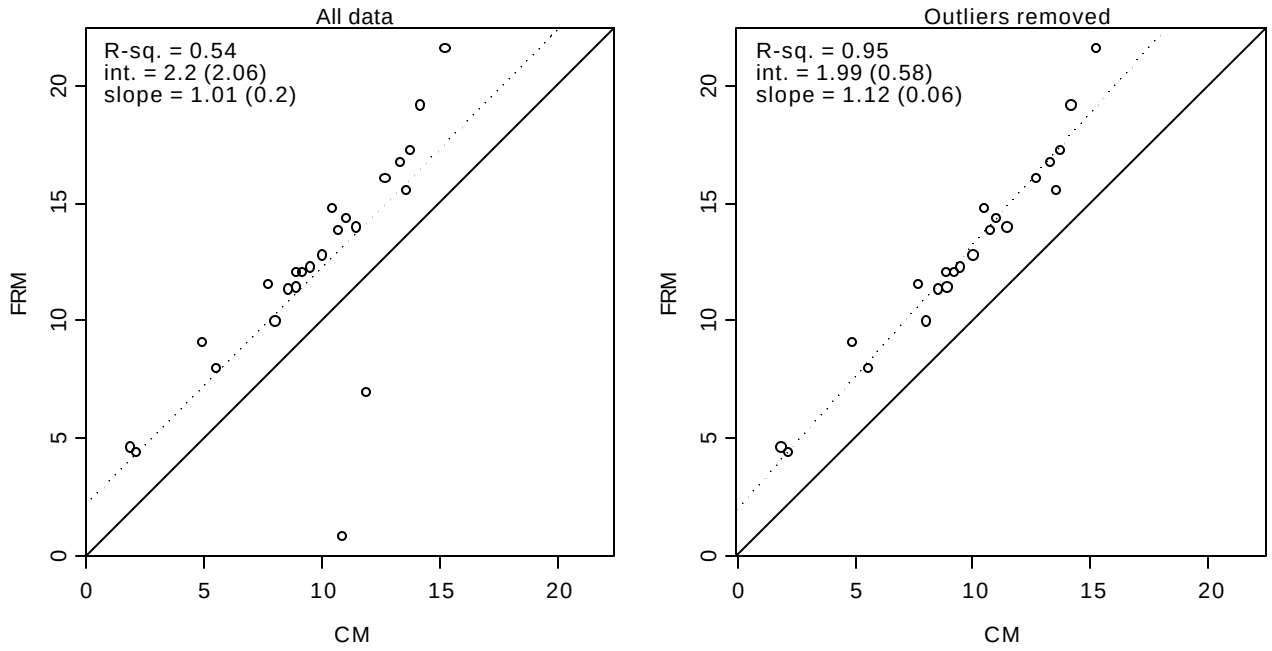

Figure 3-5. An example of the impact of outliers for Texas MSA data. The two scatter plots are before (left) and after (right) removing two outliers from the data. A regression summary is given in the upper left part of each graph.

It is also instructive to create time series plots for the monitors and overlay these so that you can see the commonalties and differences. Figure 3-6 provides an example of such a plot for the two sites in the Utah MSA with co-located continuous and FRM data. The two main purposes for the time series plots are to look for seasonal patterns and unusual time periods within the data. Seasonal or weather-related patterns might indicate that you would want a model that adjusts for these patterns. No serious anomalies appear in Figure 3-6; however, differences between continuous and FRM measurements appear to increase with concentration. This indicates a general bias between the two measurements that increases with concentration.

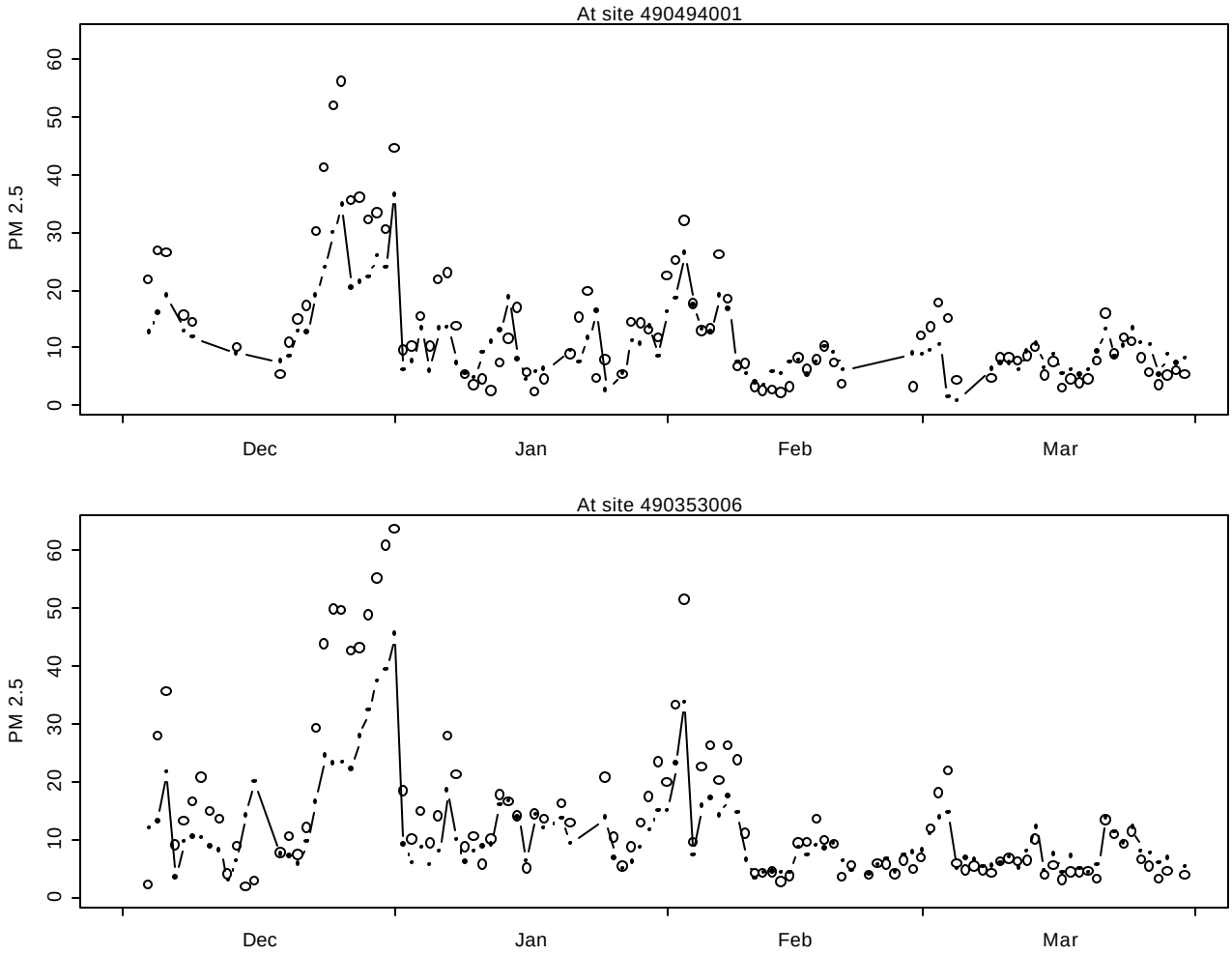

Figure 3-6. Time series of $PM_{2.5}$ concentrations at the co-located sites in the Utah MSA. The FRM measurements are circles and the continuous measurements are dots connected with a line.

As another example, Figure 3-7 shows a time series scatter plot of the difference in $PM_{2.5}$ measurements between an FRM and continuous monitor, on the natural log scale, for North Carolina MSA data (top) compared to Iowa-Illinois MSA data (bottom). Each time series in Figure 3-7 includes an overlay of a smooth trend estimate for the data. No discernible seasonal pattern is observed in the North Carolina MSA data, which is not the case for the Iowa-Illinois MSA data. A seasonally-related deviation between the continuous and FRM measurements is apparent

in the Iowa-Illinois MSA. This suggests a seasonal or weather-related adjustment may be required in this case in order to improve the modeled relationship between continuous and FRM $PM_{2.5}$ data. See the Iowa-Illinois MSA case study in Appendix B for further details on seasonal adjustment.

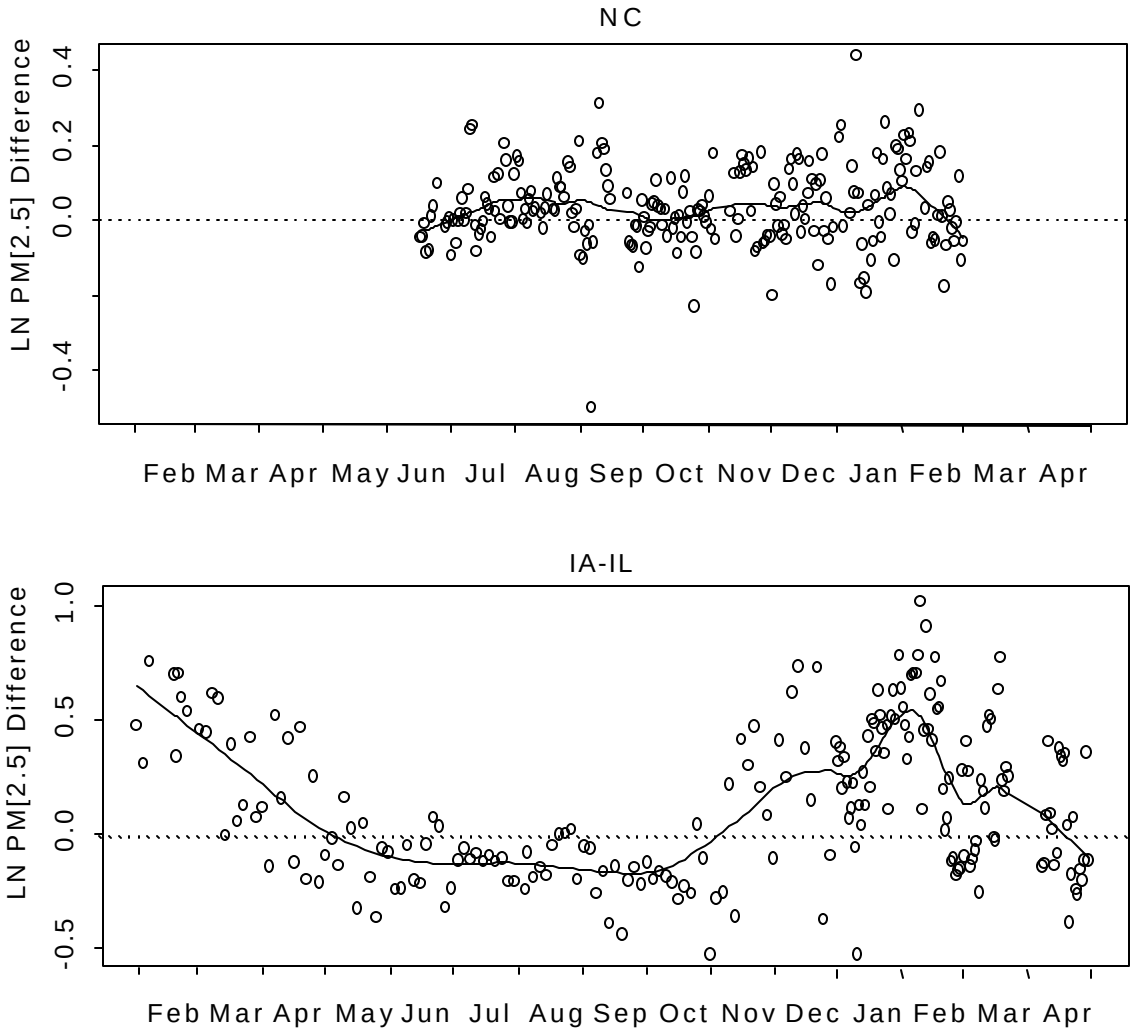

Figure 3-7. Time series, along with smooth trend, of the difference in $PM_{2.5}$ estimates on the natural log scale [i.e., ln(FRM $PM_{2.5}$) - ln(continuous $PM_{2.5}$)] for both the NC MSA (top) and the IA-IL MSA (bottom).

(Optional) Pick a continuous or an FRM monitor that has daily data and make a scatter plot of each day versus the previous day (by season) and then again comparing every third day. This is a check for something you do not want to see, autocorrelation. No relationship is good. A linear relationship indicates one of two things (in general), autocorrelation or a strong seasonal pattern. The first indicates that you may need a more complex model structure (hard to address without statistical software). The second just indicates that adding a meteorological or seasonal component to your model may be beneficial (not hard).

3.5 STEP 5 – MODEL DEVELOPMENT

We are assuming that you will, at least initially, develop separate models for each continuous-FRM pair of monitors with a sufficient number of days worth of data. This can help locate an anomalous monitor. (There are many reasons why a monitor is not in line with the others, beware of geographic barriers.) Linear regression is available with most spreadsheet packages and data plotting tools. Find the slope, intercept, and R^2 value for each pair of monitors that you are comparing or developing a relationship between. Look for slopes that are significantly different from one, intercepts that are significantly different from zero, and $R^2 > 0.80$. Slopes different from one and/or intercepts different from zero indicate a general bias between the continuous and FRM measurements. $R^2 > 0.80$ indicates a potentially good model fit.

For example, Table 3-2 summarizes regression model results for three sites in the Texas MSA, all of which provided co-located continuous and FRM measurements. Based on this initial summary, site 483390089 was eliminated from consideration for further model development. It turns out this site possessed several large outliers, which substantially degraded the regression results. However, since only 24 observations were available for model development in the first place, and since it was not clear why the observations in question were outliers, this site was eliminated from further consideration. The remaining two sites, which demonstrate reasonable model fits ($R^2 > 0.80$), both demonstrate a general bias between continuous and FRM measurements (i.e., intercepts different from zero and slopes different from one).

Table 3-2. Regression summary statistics based on comparing three sites of co-located FRM and continuous log-transformed $PM_{2.5}$ measurements in the Houston, Texas MSA

Site	N	Intercept	se(Int.)	Slope	se(Slope)	R^2	RMSE
483390089	24	1.086	0.544	0.580	0.243	0.206	0.618
482010026	82	0.915	0.086	0.714	0.039	0.811	0.147
482011039	31	0.724	0.098	0.773	0.046	0.908	0.094

3.6 STEP 6 - CONFIRMING THE RESULTS AND IDENTIFYING THE SPATIAL EXTENT OF THE RESULTS

Go back to the scatter plots of the data and add in the regression line. Are the results as you expected? Next plot R^2 versus the distance between the monitors for each pair of comparisons. Do the R^2 values follow a decreasing trend? Beware they may not be exactly decreasing, especially if you were unable to use the same days for all pairs. The drop off in the north-south direction may not be the same as the drop off in the east-west direction. Is there an FRM/continuous monitor that is significantly out of line with the others? Do the continuous monitors behave similarly? If there is a nice pattern to this plot, then you can estimate the spatial range of your model(s).

The Iowa-Illinois MSA data provide a good example of a continuous monitor performing differently than other continuous monitors. Figure 3-8 shows R^2 values (vertical axis) obtained from comparing continuous and FRM monitors, both co-located and not co-located. The horizontal axis of the plot indicates the distance in miles between the continuous and FRM monitors used in the comparison. As expected, there is a general decreasing trend in the strength of the relationship between continuous and FRM measurements as a function of increasing distance between the monitors. However, Figure 3-8 suggests that the continuous monitor 191630013 data behave somewhat differently than the data from the other two continuous monitors under study. Data from continuous monitor 191630013 yield R^2 values that fall well below the expected trend based on the other two continuous monitors, and most likely cannot be used to develop a continuous-FRM model. Figure 3-8 also reveals that R^2 values quickly fall below a level of 0.80 when data other than co-located continuous and FRM measurements are used in the Iowa-Illinois MSA model development.

3.7 STEP 7 – DECISION TIME

Do you need to go on or can you use the regression results from the previous step? This depends on how good the results were and what your needs are. Tables 2-2 and 2-3 and Appendix B can guide you in making this decision based on your R^2 value and the number of data points used to develop the model.

For example, in the case of the North Carolina MSA, the continuous-FRM relationship is so strong that a very good model is achieved, using only simple linear regression, for virtually all pairs of continuous and FRM comparisons. In the case of the Iowa-Illinois MSA, a little more work is required to develop a seasonal or meteorological adjustment that improves the model to the point of acceptance. For the Texas MSA, one site of co-located continuous and FRM data yields a strong relationship with little effort, another co-located site produces a marginally adequate model that might require improvement, and a third co-located site lacks sufficient data and model adequacy for further development. Finally, without log-transforming the Utah MSA data, potentially acceptable models are achieved (i.e., $R^2 > 0.80$) at the two sites of co-located FRM and continuous data. Further inspection of the Utah MSA data suggests that model improvements might be obtained by carefully considering the effect of several observations that appear as potential outliers.

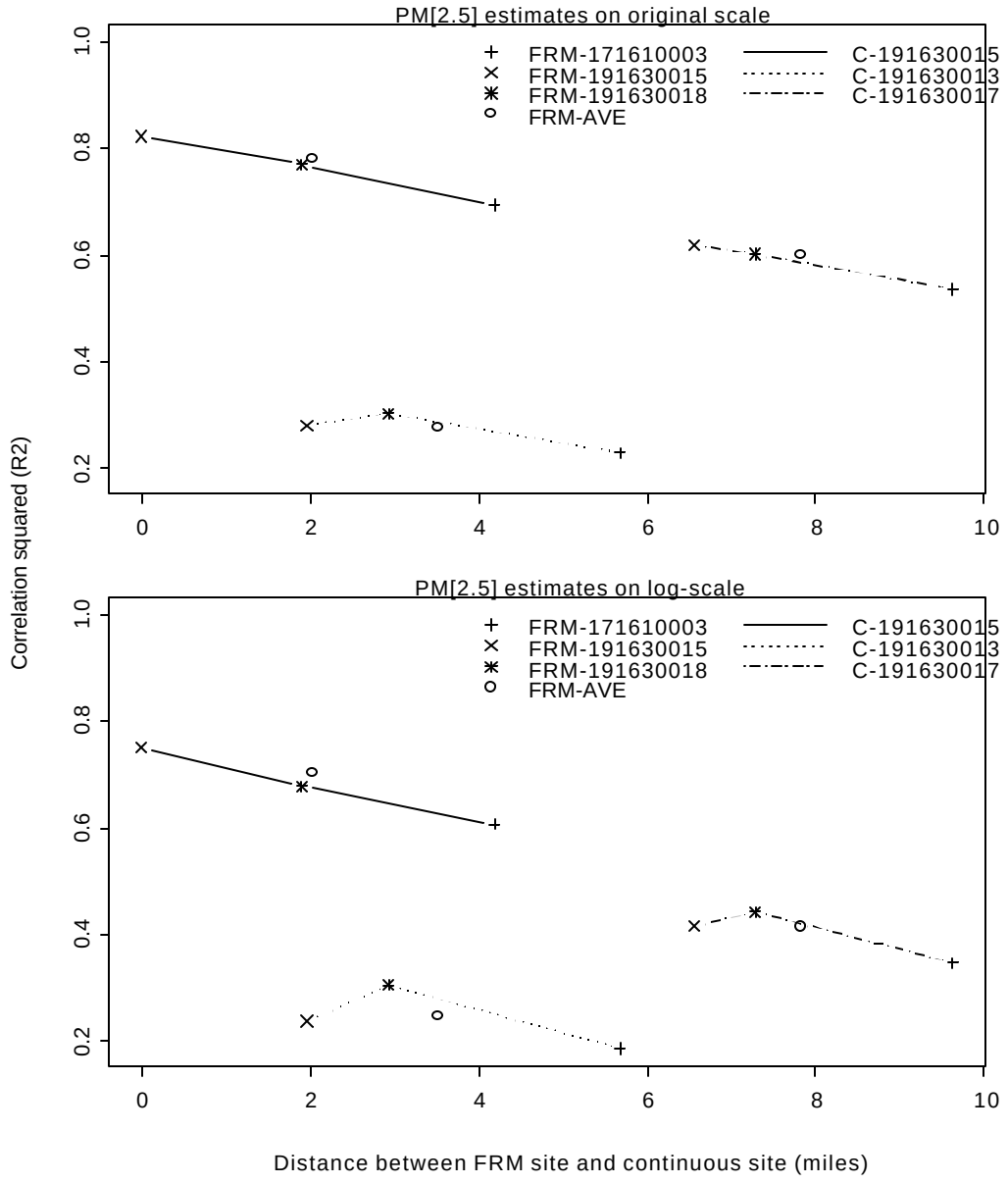

Figure 3-8. **R^2 values between different FRM monitors (different symbols) and continuous monitors (different line types) plotted versus the distance between the sites, based on Iowa-Illinois MSA data. The two graphs correspond to $PM_{2.5}$ estimates on the original scale (top) and on the log-transformed scale (bottom).**

3.8 STEP 8 – IMPROVING THE MODEL WITH AUXILIARY DATA

If the linear model based on the continuous monitor data alone do not meet your needs, then there are a variety of sources of auxiliary data that can be used to improve the R^2 value. For a simple example, suppose the PM composition changes from season to season. This may cause the relative response of the continuous monitor to change. Simply allowing for different slopes and / or intercepts for each quarter may be sufficient to improve the overall fit between the two. Other types of seasonal adjustments might work as well. For example, at the Iowa-Illinois MSA site with co-located FRM and continuous data, the R^2 value improved from 0.693 to 0.840 when including a sinusoidal seasonal adjustment (i.e., a smooth, periodically recurring, seasonal trend) in the model (see Appendix B).

There are many other possibilities that you could include, such as adjustments for using meteorological data: wind direction and speed (if, for example, your main source of PM is from the north, you can use this information), barometric pressure, mixing height, temperature, etc. For example, at the Iowa-Illinois MSA site with co-located continuous and FRM data, the R^2 value improved from 0.693 to 0.856 when including an adjustment for temperature (i.e., daily average temperature) in the model (see Appendix B). Note that this adjustment was a slight improvement over the seasonal adjustment considered for these same data.

There is no single right answer. Keep trying until you get a model fit that meets your need. Chances are that the more variability you have in the chemical composition of the PM and in the atmospheric conditions of your region, the more adjustments you will need.

3.9 STEP 9 – FINAL CHECKS

If you electronically report your continuous monitor results, for example to a webpage, then make sure that the model is incorporated appropriately [e.g., untransform (exponentiate) your model results if you use a log-transform]. For example, consider the model developed for the North Carolina MSA site with co-located FRM and continuous log-transformed data, which concludes:

$$\ln(\text{FRM}) = -0.114 + 1.054 * \ln(\text{continuous}) .$$

Suppose a continuous $PM_{2.5}$ measurement of 20 $\mu g/m^3$ were observed. Then the appropriate model-based FRM value (i.e., the FRM value based on continuous data calibrated according to the model) to use in reporting the AQI would be:

$$FRM_{model} = \exp\{[-0.114+1.054*\ln(20)]\} = \exp\{3.0435\} = 20.98\ \mu g/m^3.$$

Plugging 20.98 $\mu g/m^3$ into the formula for the AQI yields a reported index value of:

$$PM_{2.5}: \quad \frac{(100-51)}{(40.4-15.5)}(20.98-15.5)+51=62$$

In summary, the resulting AQI value is derived from a modeled FRM measurement, where the modeled FRM measurement is based on a continuous $PM_{2.5}$ measurement and the model relating continuous and FRM measurements.

Finally, how often you check and update your model depends on how varied the monitors in your area tend to be. It will probably take at least a quarter's worth of data to make any significant change, unless you have made changes in the operating procedures of your continuous monitor. Also, if you have used seasonal adjustments or parameters that change significantly from season to season, then quarterly checks are probably warranted.

APPENDIX A:

STATISTICAL ASSUMPTIONS UNDERLYING DQO TABLES 2-2 AND 2-3

February 2001

APPENDIX A: STATISTICAL ASSUMPTIONS UNDERLYING DQO TABLES 2-2 AND 2-3

The statistical parameter R^2 has been defined as the parameter of interest for determining whether the model relating FRM with continuous $PM_{2.5}$ measurements is acceptable. This appendix provides details regarding the statistical assumptions for R^2 that were used to derive Tables 2-2 and 2-3 in Section 2.7 of this report.

As stated in Section 2.6, in simple or multiple regression, R^2 is the square of the correlation coefficient between observed FRM $PM_{2.5}$ data values and their associated modeled values derived from a fitted statistical linear model. This interpretation is the basis for establishing the R^2 distributional assumption. First, we define the statistic W as follows:

$$W = \frac{1}{2}\ln\left(\frac{1+R}{1-R}\right).$$

Assuming the observed FRM $PM_{2.5}$ data values and their associated predictions from the model follow a bivariate normal distribution, it follows that W has an approximate normal distribution with mean ½ ln [(1 + D) / (1 - D)] and variance $1/\sqrt{n-3}$, where D equals the square root of the true unknown R^2 value. Testing a null hypothesis of H_0: $R^2 \# 0.60$ is thus equivalent to a test of

$$H_0: \; W \le \frac{1}{2}\ln\left(\frac{1+\sqrt{0.60}}{1-\sqrt{0.60}}\right) = 1.0317 .$$

To conduct an "-level test (i.e., require false rejection decision error to be below "), we require

$$P_{R^2 \le 0.60}\{W > c\} \le \alpha ,$$

where the bound c is chosen to satisfy the above inequality. The bound on c will be obtained at the boundary conditions of " and R^2=0.60. Thus, we must solve for c satisfying

$$P_{R^2 = 0.60}\{W > c\} = \quad .$$

This is equivalent to

$$P\left\{Z > (c - 1.0317)(n-3)^{1/4}\right\} = \alpha \ .$$

where W has been transformed to Z (a standard normal random variable with a mean of zero and a variance of one) by subtracting off the mean of W when R^2=0.60 (1.0317) and dividing by the standard deviation of W $\left(\left[1/\sqrt{n-3}\right]^{1/2}\right)$. Based on the distribution of Z, this equality is satisfied when

$$(c - 1.0317)(n-3)^{1/4} = z_{\alpha},$$

where $z_{"}$ is the "th percentile of the standard normal distribution. Solving for c gives

$$c = \frac{z_{\alpha}}{(n-3)^{1/4}} + 1.3017.$$

Next, to obtain sample size requirements, we consider the DQO parameters $ and). Our requirement is

$$P_{R^2 \geq 0.60+\Delta}\{W > c\} \geq (1-b).$$

This is equivalent to

$$P\left\{ Z > \left[c - \frac{1}{2}\ln\left(\frac{1+\sqrt{0.60+\Delta}}{1-\sqrt{0.60+\Delta}} \right) \right](n-3)^{1/4} \right\} \geq 1-b\ ,$$

where, as above, W has been transformed to Z by subtracting its mean and standard deviation assuming $R^2 = 0.60 +)$. Based on the distribution of Z, this equality is satisfied when

$$\left[c - \frac{1}{2}\ln\left(\frac{1+\sqrt{0.60+\Delta}}{1-\sqrt{0.60+\Delta}} \right) \right](n-3)^{1/4} \leq z_{1-b}\ ,$$

where $z_{1-\$}$ is the $(1-\$)^{th}$ percentile of the standard normal distribution.

Substituting in for c and solving for n gives the formula for calculating the sample sizes of Table 2-2 in Section 2.7 as

$$n \geq \left[\frac{z_a - z_{1-b}}{\frac{1}{2}\ln\left(\frac{1+\sqrt{0.60+\Delta}}{1-\sqrt{0.60+\Delta}} \right) - 1.0317} \right]^4 + 3.$$

Finally, since c is the point at which the model is determined to be acceptable on the scale of W, we simply need to transform c back to the scale of R^2 to obtain the formula for calculating the R^2 lower bounds of Table 2-3 in Section 2.7 as

$$R^2 \geq \left[\frac{\exp(2c) - 1}{\exp(2c) + 1} \right]^2,$$

where, given the specification of n above, c is as defined previously.

Reference

Hogg, R., and Tanis, E. (1977). Probability and Statical Inference, Macmillan Publishing Company, Inc., New York, New York.

APPENDIX B:

FOUR CASE STUDIES

February 2001

Appendix B: Table of Contents

List of Tables

List of Figures

APPENDIX B: FOUR CASE STUDIES

B.1 GREENSBORO–WINSTON-SALEM–HIGH POINT, NORTH CAROLINA

The data from North Carolina (NC) has one continuous monitor (CM) at site 370670022 (in Winston-Salem in Forsyth County). The available data consist of 259 daily $PM_{2.5}$ measurements, the first one on 06/16/1999 and the last one on 02/29/2000. The CM site has a co-located federal reference method (FRM) monitor with a total of 231 $PM_{2.5}$ estimates from 06/16/1999 to 02/29/2000. In addition, there are five other FRMs nearby. Figure B-1 shows the location of the sites.

Analysis of Co-located Site

The co-located site has a total of 227 days with observations from both the FRM and the CM. An initial, exploratory, analysis is given in Figure B-2 (time series) and B-3 (scatter plots). The scatter plots of FRM versus CM measurements are done for untransformed and log-transformed data. The scatter plot for the untransformed data shows no serious outliers or influential points, although, one point in the upper-right corner shows larger deviation from the 45-degree line than surrounding observations. A summary of the least squares regression fits are given in Table B-1.

Table B-1. Summary of Least Squares Regression Results when Regressing FRM Versus CM at the Co-Located Site in NC.

	N	Intercept	se(Int.)	Slope	se(slope)	RMSE
untransformed	227	0.026	0.232	1.040	0.013	1.595
log-transformed	227	-0.114	0.036	1.054	0.013	0.104

The summary in Table B-1 indicates a very strong relationship between the FRM and the CM measurements. A diagnostic is given in Figures B-4 (histograms of residuals), B-5 (scatter plot of

residuals versus predicted), and B-6 (time series of residuals). None of the diagnostics reveal serious problems. Using untransformed data results in a histogram of the residuals slightly skewed to the right (Figure B-4, right) and a residuals spread that increases with larger $PM_{2.5}$ values (Figure B-5, right), which is not evident when log-transforming the data. Thus, from a statistical point of view, the log-transformed regression deviates less from the normal assumption underlying the regression (the histogram of the residuals is not skewed and the spread of the residuals does not depend on the size of the predicted value). Finally, the residuals from the least square regressions do not show any seasonal trend (Figure B-6), hence, no seasonal adjustment is needed.

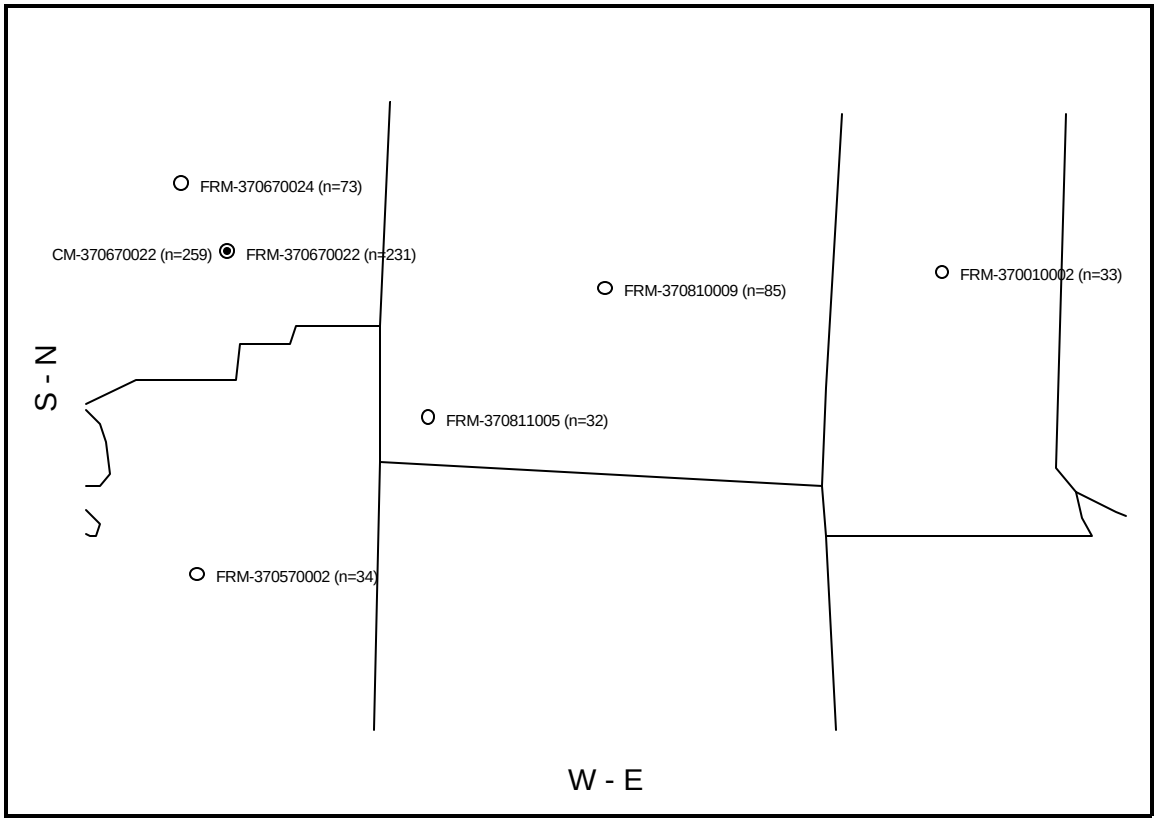

Figure B-1. Available sites in NC (circles are FRM sites, black dots are CM sites). The number shown in parentheses is the number of observations available from 06/16/1999 to 02/29/2000.

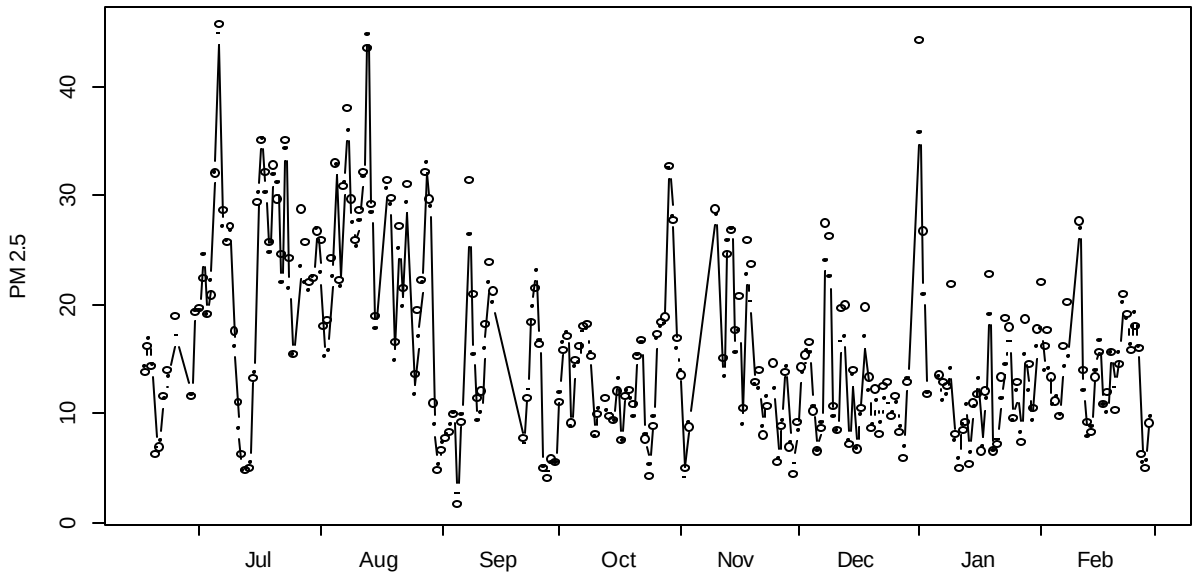

Figure B-2. **Time series of PM 2.5 daily estimates at the co-located site in NC. Circles are FRM estimates and black dots, connected with solid line, are CM estimates.**

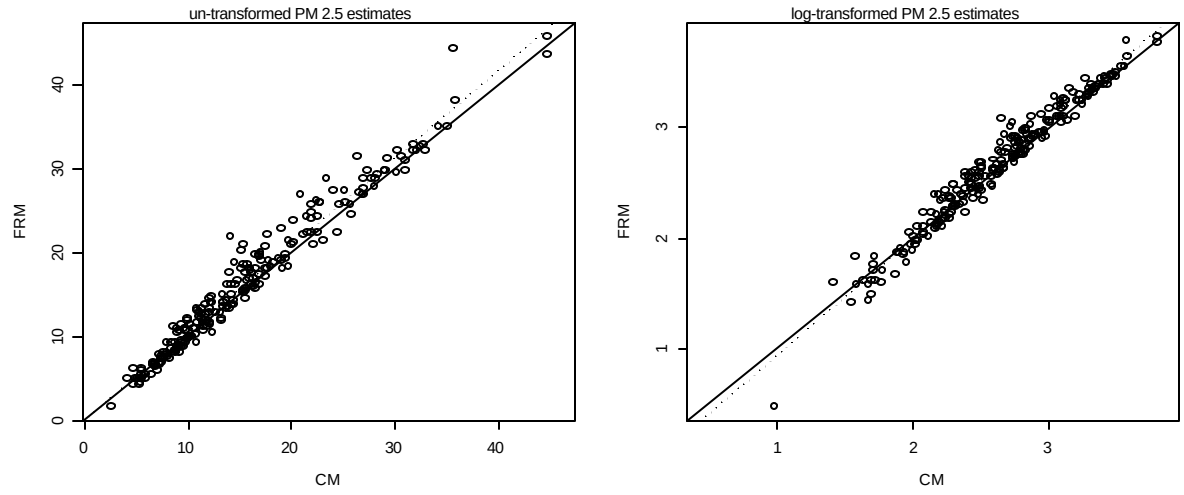

Figure B-3. **Scatter plot of FRM PM 2.5 daily estimates versus CM PM 2.5 daily estimates for untransformed data (left) and log-transformed (right). The solid line is the 45 degree line and the dotted line is a least square regression line.**

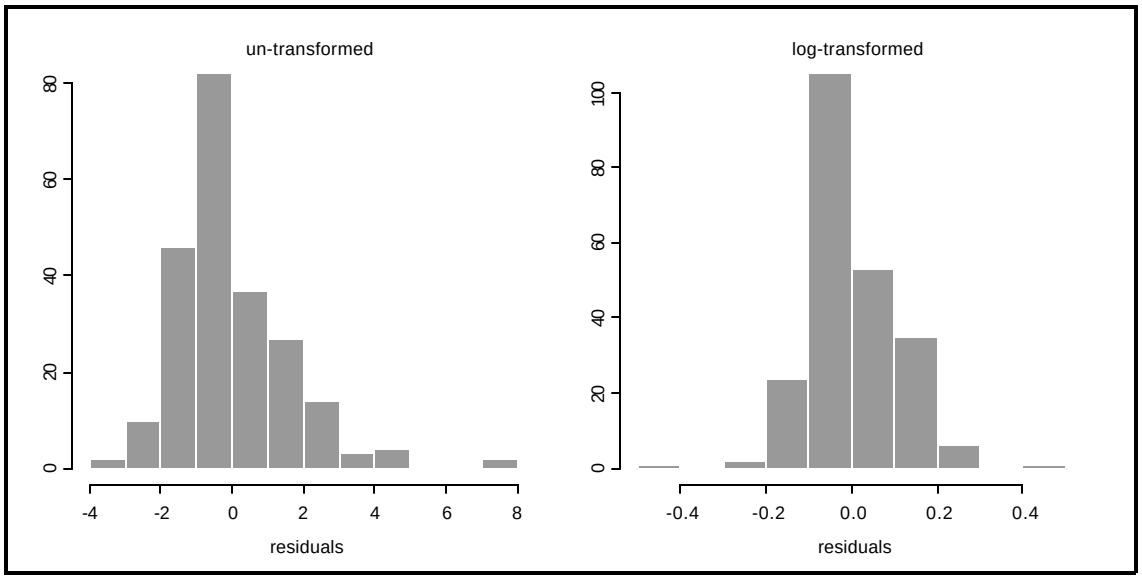

Figure B-4. **Histogram of the residuals from the least squares regression of FRM versus CM for untransformed $PM_{2.5}$ estimates (left) and log-transformed (right)**

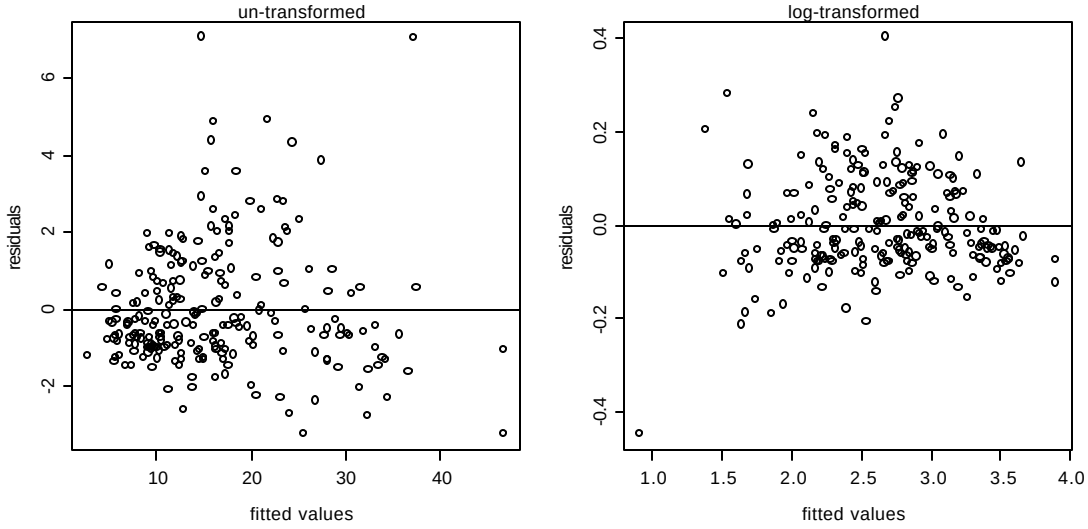

Figure B-5. **The residuals from the least squares regression of FRM versus CM plotted versus the fitted values from the same regression for untransformed $PM_{2.5}$ estimates (left) and log-transformed estimates**

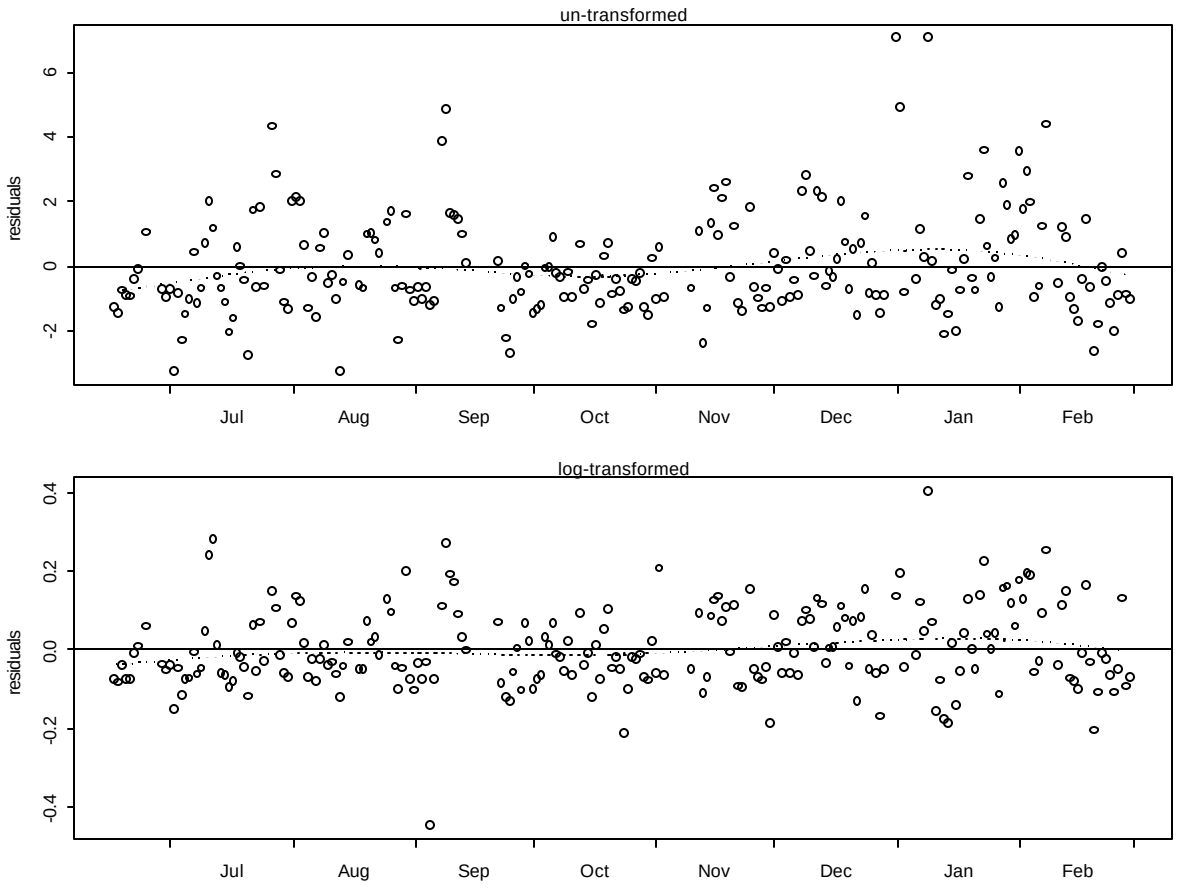

Figure B-6. The residuals from the least squares regression of FRM versus CM plotted versus time for the untransformed data (top) and log-transformed data (bottom)

Analysis of Other Available Data

Given the strength of the CM-FRM relationship observed at the co-located site in the NC MSA, it may be worth considering a comparison of CM and FRM data that are not co-located. As such, a study was conducted to determine how the strength of the CM-FRM relationship observed for co-located data might change when the comparison is made using data not co-located. If the relationship remains strong, an MSA might consider developing multiple CM-FRM models, or develop

a model using a combination of all or most of the available FRM data in the MSA. For example, develop a model for the average of the set of daily FRM measurements with CM measurements, which yields transformed CM measurements that are more representative of the overall MSA spatial region. Or, if the MSA regularly uses each FRM monitor to calculate a set of AQI's, then develop a separate model between the continuous monitor and each FRM. This would give the MSA the ability to report an analogous set of continuous-based AQI's.

There are only thirteen days where all six FRMs and the CM have daily PM estimates, but by ignoring the FRM at site 370811005 and using the remaining five FRMs, then there are eighteen days with estimates from all monitors. Figure B-7 shows the time series of the five FRMs and the CM, and Figure B-8 shows the scatter plots of the five FRMs versus the CM. The scatter plots show no problematic observations and indicate a good correlation between all five FRMs and the CM. Table B-2 confirms the strength of the correlation in a regression summary table, based on log-transformed data. In addition to regressing the five FRMs versus the CM, the average of the five FRMs was also used (the bottom line of the table). Table B-2 also shows how the correlation decreases slowly with increasing distance between the monitors. This is better seen in Figure B-9, which shows R-squared versus distance for both untransformed data and log-transformed data.

Table B-2. Least square regression summary for each of the five FRMs versus the CM. The last column shows the distance (miles) between the monitors.

FRM	n	intercept	se(int)	slope	se(slope)	R-squared	Distance
370670022	18	-0.100	0.072	1.032	0.024	0.991	0.0
370670024	18	-0.102	0.087	1.015	0.029	0.987	5.3
370570002	18	0.212	0.167	0.947	0.056	0.948	20.7
370810009	18	-0.017	0.151	1.016	0.051	0.962	24.3
370010002	18	0.120	0.197	0.968	0.066	0.932	45.7
Average	18	0.039	0.102	0.991	0.034	0.982	19.2

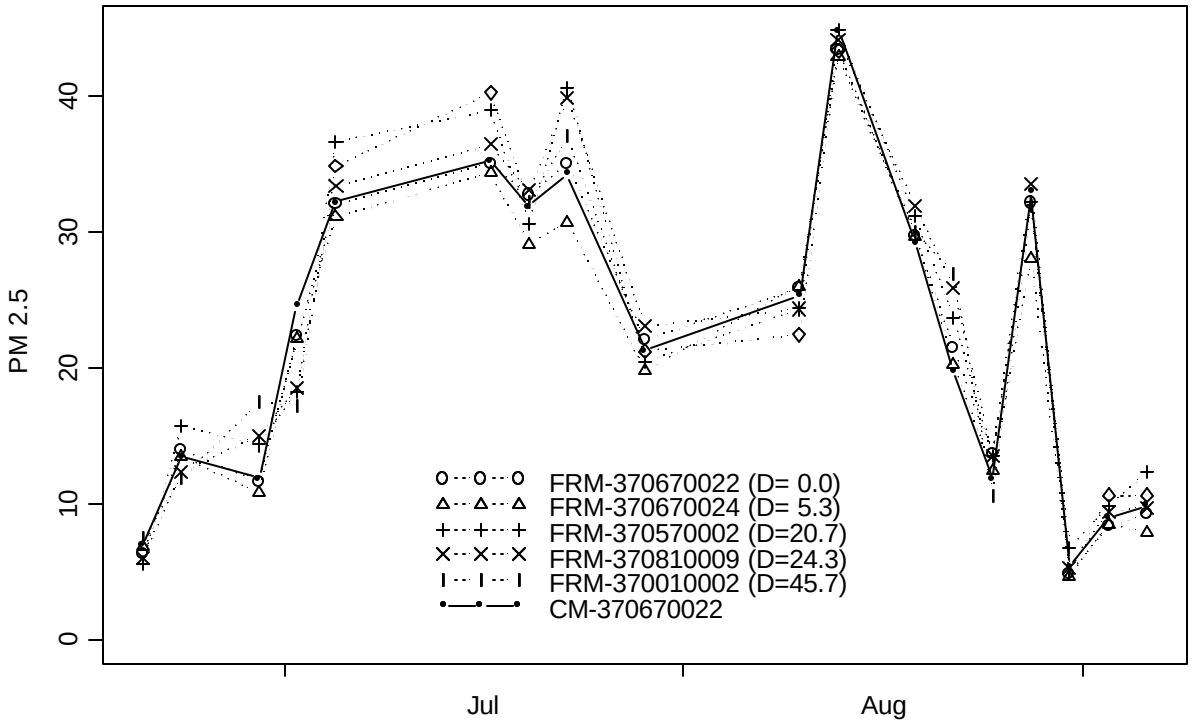

Figure B-7. Time series of daily $PM_{2.5}$ concentrations from five FRMs and one CM in NC

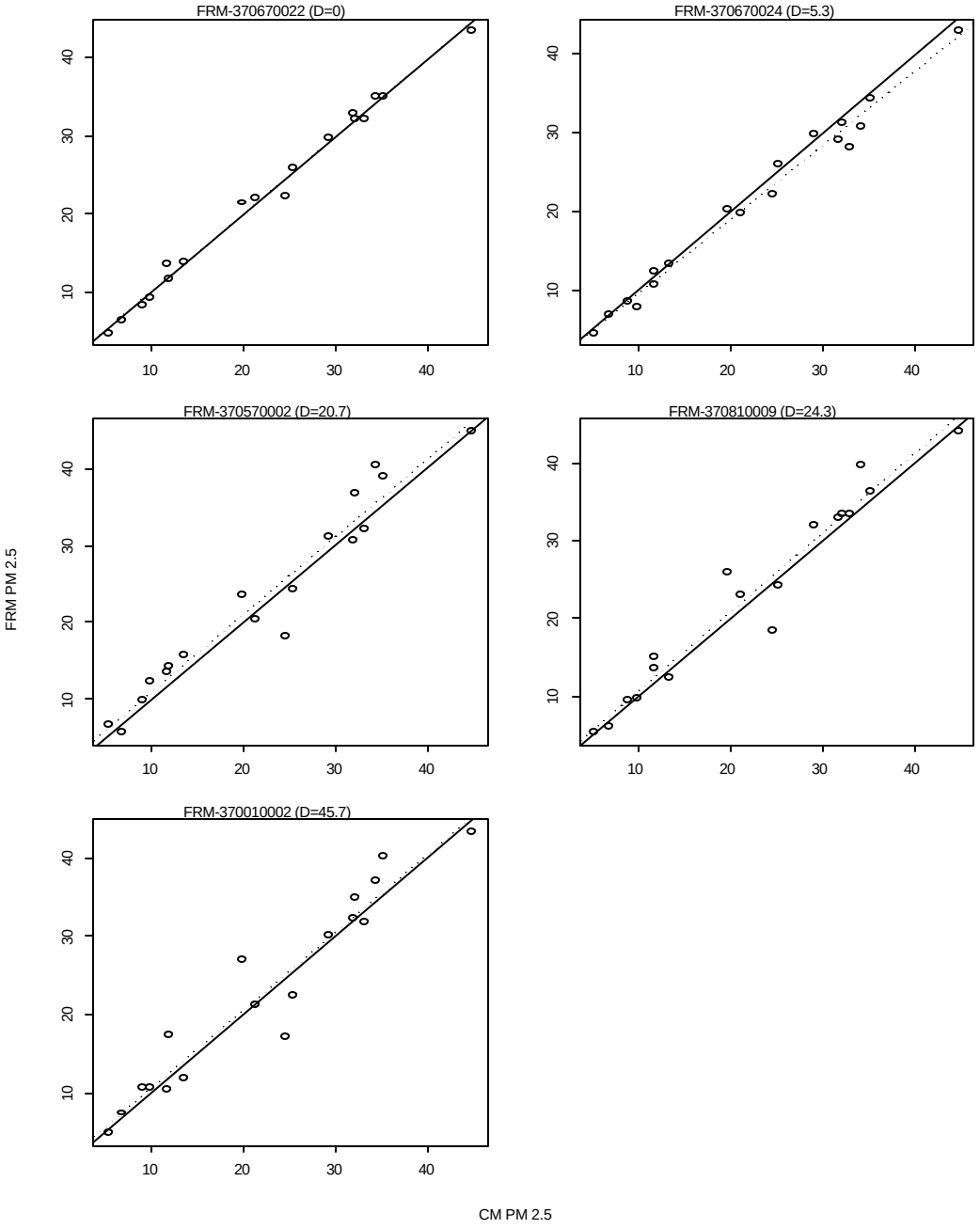

Figure B-8. Scatter plot of each of the five FRMs versus the CM. Also shown is the 45-degree line (solid) and the least square fit (dotted line)

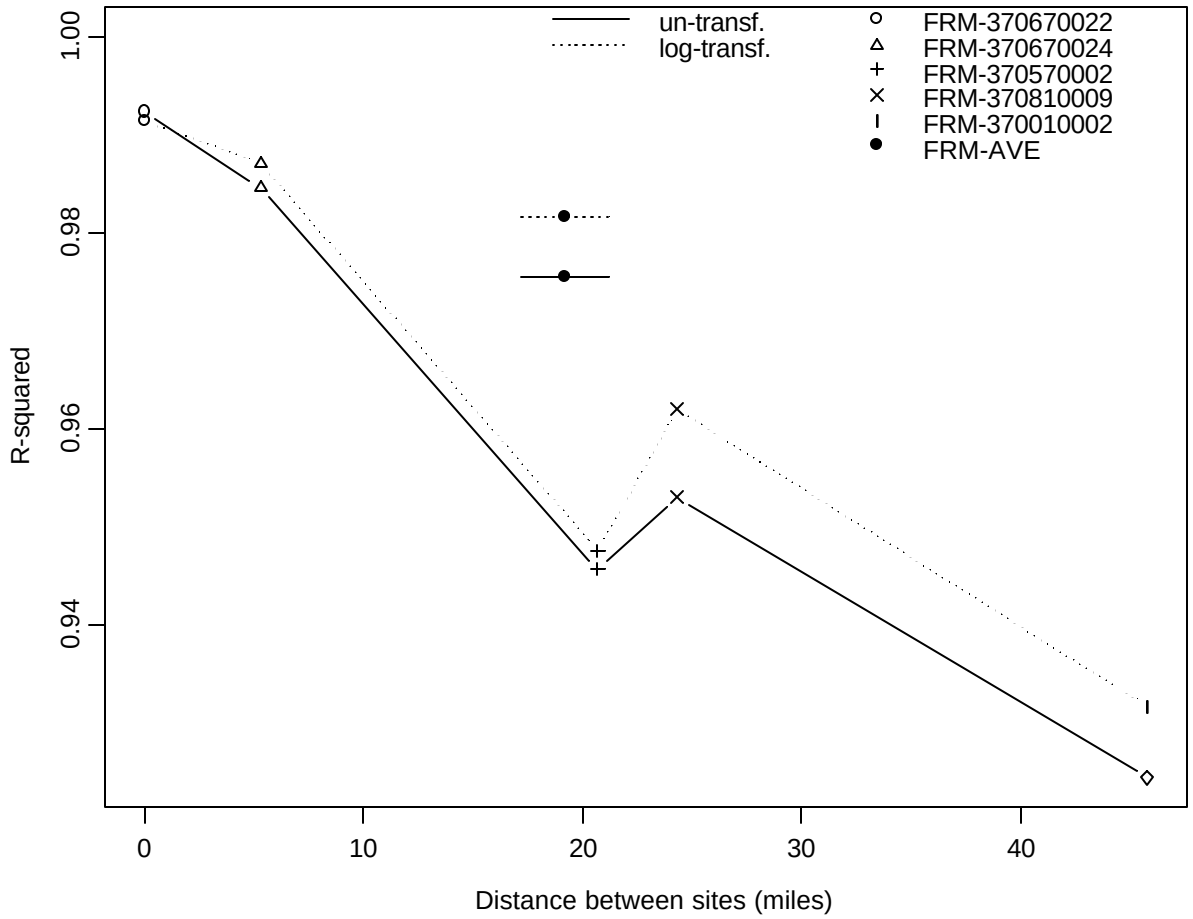

Figure B-9. R-squared between the FRMs (and their average) and the CM plotted versus the distance between the monitors

Conclusions

The strength of the CM-FRM relationship appears strong at the NC MSA, whether the monitors used in the comparison are co-located or not. This leaves several options for using FRM data in developing a model, all of which appear reasonable. Ideally, a log-transformation of the data would be made before developing a model. Results for log-transformed data appear somewhat better. In particular, common regression model assumptions such as constant variability across observations and symmetrically distributed errors appear to be more closely satisfied under the log-transform. However,

in the interest of simplicity, models based on untransformed data appear adequate as well. The choice of whether or not to transform the data depends on the level of complexity the MSA might want to introduce into the model development process. Similarly, the choice of whether to use FRM data other than the co-located site to develop model(s) will depend on the amount of data analysis and model development the MSA wants to pursue.

B.2 DAVENPORT-MOLINE-ROCK ISLAND, IOWA-ILLINOIS

The IA-IL MSA data has three CMs and three FRMs, but only one co-located site (191630015). Two of the CMs are sampled from 01/01/1999 to 04/30/2000, and the third one from 02/01/1999 to 04/30/2000. The co-located FRM is sampled from 02/27/1999 to 04/30/2000, with sampling frequencies of every third day in 1999 and every day in 2000. The other two FRMs are sampled from 07/02/1999 to 04/30/2000 (approximately every third day) and from 01/06/1999 to 03/31/2000 (approximately every sixth day). See Figure B-10 for the location of sites and number of observations available. Based on the available data, the co-located CM can be calibrated using the FRM at that site, but the other two CMs need to be calibrated using FRM data (or the average of several FRMs) at nearby sites.

Analysis of Co-located Site

There are 214 days with $PM_{2.5}$ daily estimates from both the FRM and the CM at site 191530015. Figure B-11 shows the two time series and Figure B-12 shows two scatter plots, one for untransformed data and one for log-transformed data. From these two figures it is evident that there is not good correlation between the two monitors (the scatter plots in Figure B-12). The time series plot shows also that there is much better correspondence between the two monitors in the summer, but in the winter time the CM reports, in general, higher $PM_{2.5}$ concentrations that the FRM (Figure B-11). Analysis of the residuals from the least squares

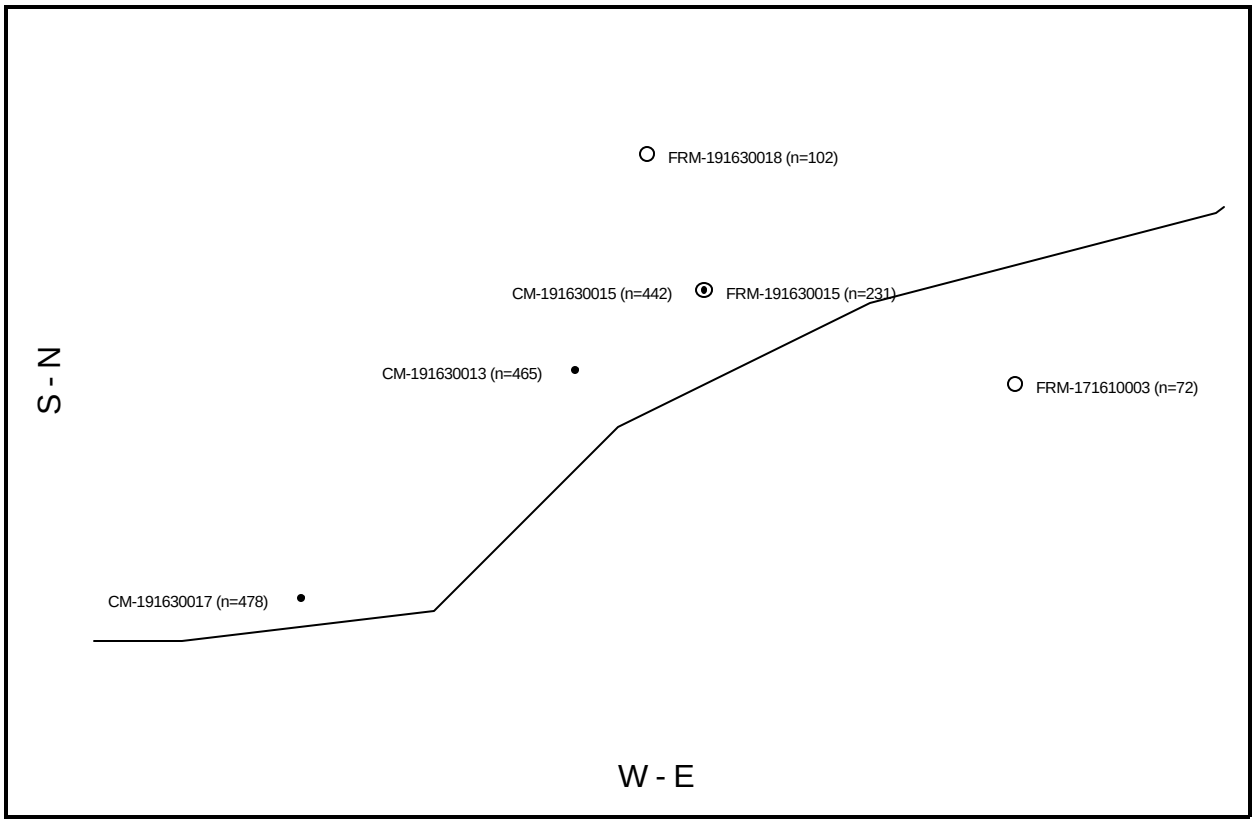

Figure B-10. Location of sites in the IA-IL MSA (circles are FRM sites, dots are CM sites). The number of observations available is shown in parenthesis.

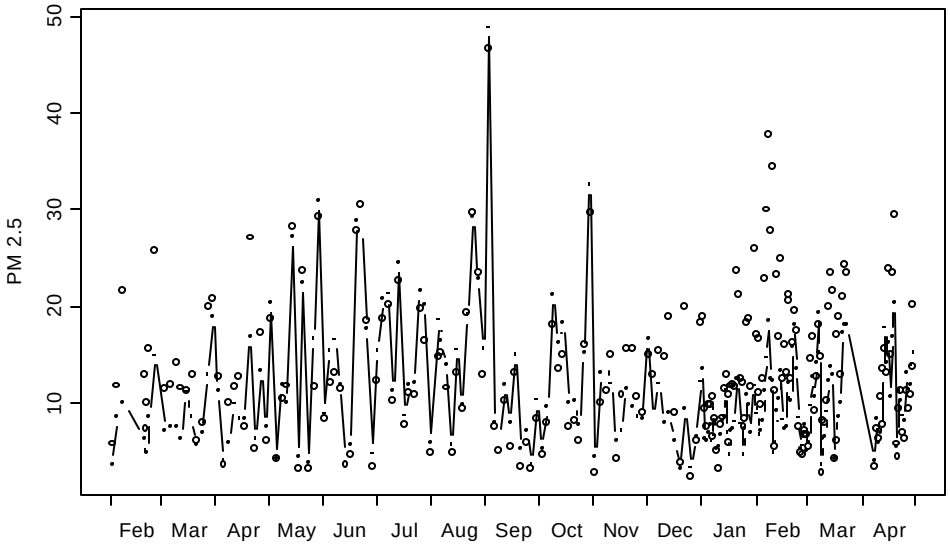

Figure B-11. Time series of $PM_{2.5}$ measurements at the co-located site in the Iowa-Illinois MSA. Circles are FRM estimates and connected black dots are CM estimates

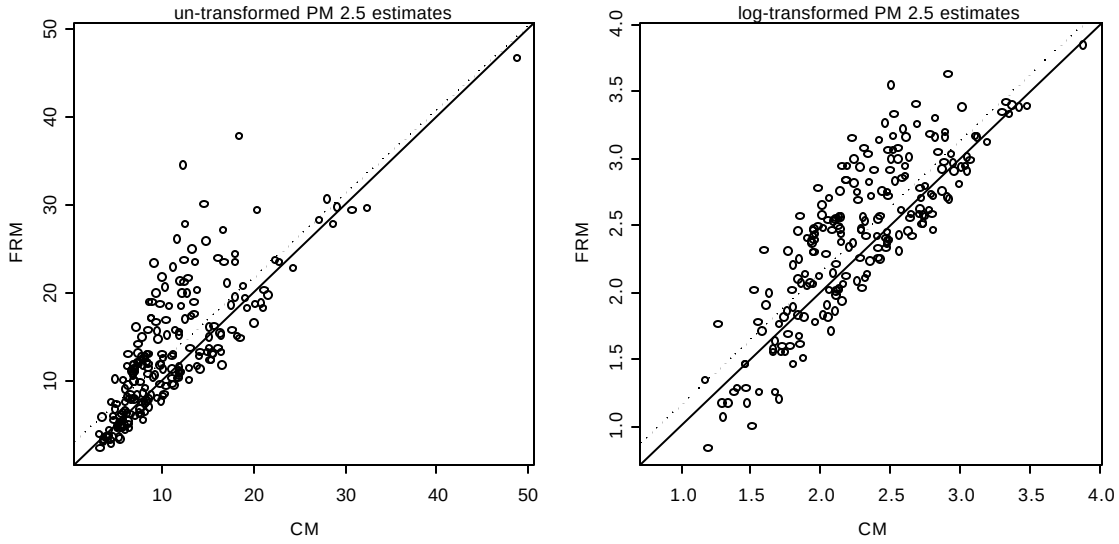

Figure B-12. Scatter plot of FRM $PM_{2.5}$ values versus CM $PM_{2.5}$ values for untransformed data (left) and log-transformed (right). The solid line is the 45 degree line and the dotted line is a least squares regression line

regression shows that log-transforming the data seems to be appropriate (Figure B-13 and Figure B-14). Figure B-11, reveals the seasonal behavior in the data, namely, the CM underestimates the FRM in the winter. The results from a least squares regression can be seen in Table B-3.

Table B-3. Summary of least squares regression results when regressing FRM versus CM at a co-located site.

	N	Intercept	se(Int.)	Slope	se(slope)	R^2	RMSE
untransformed	214	2.661	0.638	0.956	0.050	0.631	4.512
log-transformed	214	0.173	0.106	0.988	0.045	0.693	0.327

Our first attempt to increase the quality of the model is to include a smooth, periodic, seasonal trend in the model. More precisely, let *d* denote the day number within the year, and Y_d and X_d the $PM_{2.5}$ estimates from the FRM and CM, respectively, from that day. Then the basic regression model, on the log-scale, is:

$$\log(Y_d) = a + b \log(X_d) + e_d,$$

where $_d$ are measurement errors, assumed to be independent and normally distributed with mean zero and standard deviation s . The basic model can be extended by adding a smooth, periodic sinusoidal seasonal trend to it. The simplest case is a sinusoidal seasonal trend with two terms:

$$\log(Y_d) = a + b \log(X_d) + g_1 \sin(d \cdot 2p / 365) + g_2 \cos(d \cdot 2p / 365) + e_d.$$

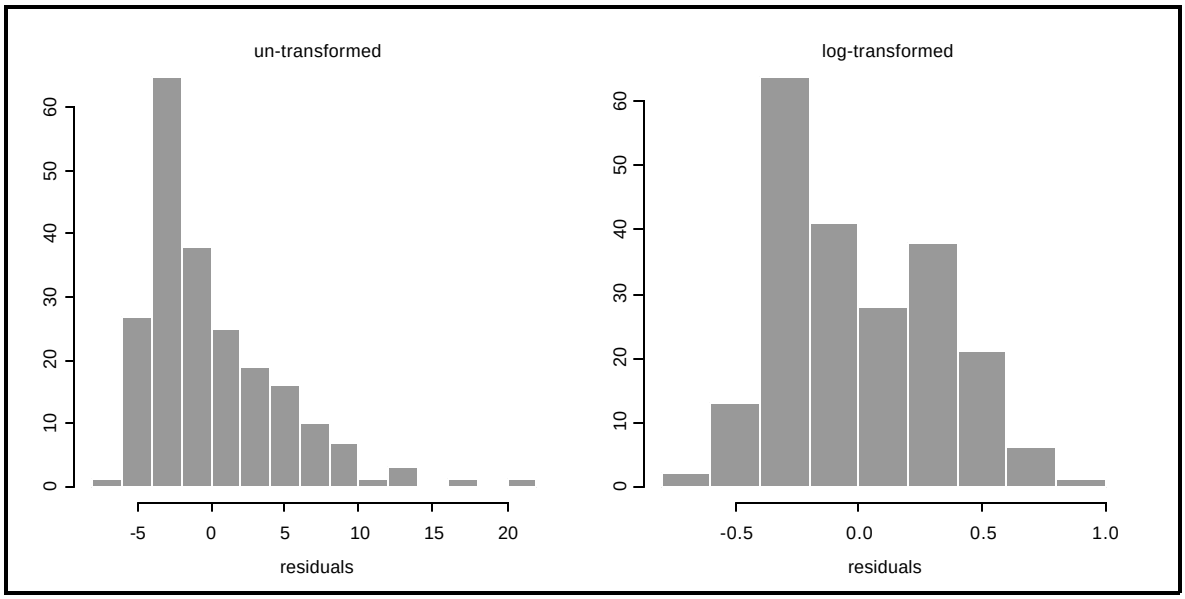

Figure B-13. Histogram of the residuals from the least squares regression of FRM versus CM for untransformed $PM_{2.5}$ estimates (left) and log-transformed (right)

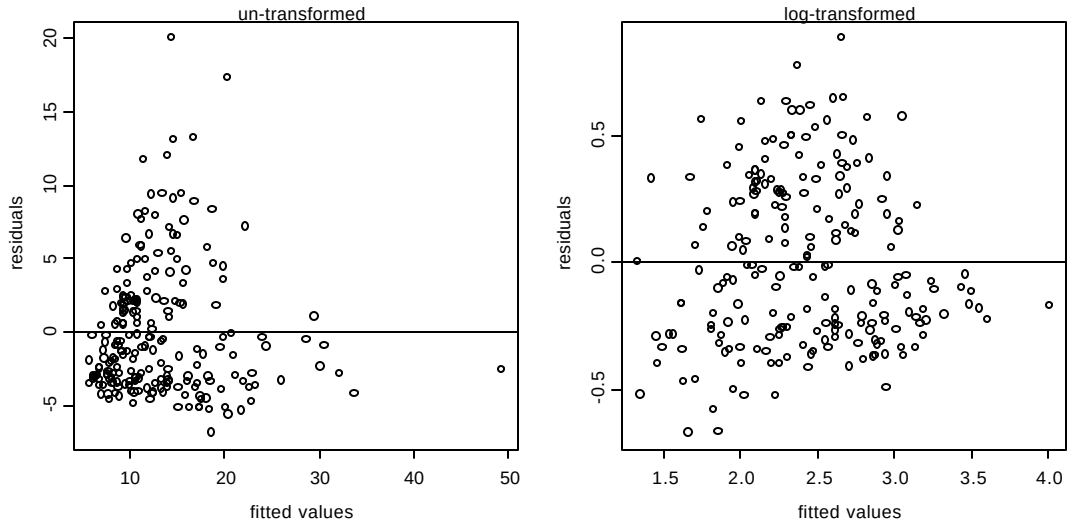

Figure B-14. The residuals from the least squares regression of FRM versus CM plotted versus the fitted values from the same regression for untransformed $PM_{2.5}$ data (left) and log-transformed (right)

It is possible to use a trend with four terms (i.e., in addition to the two terms shown above, an additional two terms are added, which are identical to the two first two but with *d* replaced with *2d*).

Three different seasonal trends were added to the basic model, a periodic sinusoidal trend with two, four and six terms. The model with four terms was significantly better than the model with two terms (p-value < 0.01), but the model with six terms was not a significant (p-value > 0.5) addition to the model with four terms. By adding the four terms seasonal trend, R^2 improved from 0.693 (the basic model with log-transformed data) to 0.840, which is acceptable according to Tables 2-2 and 2-3 of Chapter 2.

Another approach, other than seasonal adjustment, is to use meteorological data to improve the model. Daily average temperatures are available at the co-located site. The following model was applied to the data:

$$\log(Y_d) = a + b \log(X_d) + gT_d + e_d,$$

where T_d denotes the daily average temperature in day *d*. This temperature adjusted model yielded an R^2 of 0.856, slightly better than the model with a general, smooth, seasonal trend. In summary, the above discussion summarizes two approaches to improving the model for co-located Iowa-Illinois MSA data, namely adding a seasonal adjustment or incorporating a meteorological adjustment. In this case, both methods appear to improve the model to the point of acceptance.

Analysis of Other Available Data

The co-located CM could be calibrated, using a temperature adjustment, to the co-located FRM. But, this is not the case for the other two CMs, since they do not have co-located FRMs (see Figure B-10). It is therefore of importance to see if other FRMs, at nearby sites, can be used to calibrate these monitors. The first step in such an analysis is to explore the spatial variation in $PM_{2.5}$ concentrations.

There are 35 days when all six monitors (3 CMs and 3 FRMs) have $PM_{2.5}$ estimates, starting in July 1999 and lasting through January 2000. Figure B-15 compares the time series of the monitors and Figures B-16 and B-17 show the scatter plots (each FRM plotted versus each CM). Tables B-4 and B-5 summarize the least squares regressions shown in the scatter plots.

The two CMs at sites 161930013 and 191630017 do not show high correlation to nearby FRMs (Tables B-4 and B-5). Figure B-18 shows R^2 plotted versus distance between the FRMs and the CMs. Both the time series plots (Figure B-15) and the scatter plots (Figures B-16 and B-17) show why; there are days that have large deviations between FRMs and CMS, and it is not so obvious to conclude that these days are outliers (i.e., bad CM observations). In addition, we saw at the co-located site that there is a significant seasonal pattern in the deviation between the FRM and the CM. This same seasonal pattern can also be seen for sites not co-located (Figure B-15), but not at the same strength as was observed for the co-located site.

Given the relatively low R^2 values observed in Tables B-4 and B-5, an attempt was made to improve upon the basic models. First, a seasonal adjustment (two term sinusoidal seasonal trend) was added to the basic model, on the log-scale, for each of the FRM versus CM comparisons. Next, outliers from the seasonal regression model were removed. An observation was determined an outlier if its residual was larger than 2.5 times the estimated root mean squared error (RMSE) from the

seasonal regression. The CM at site 191630013 still did not produce an acceptable R^2, and the CM at site 191630017 yielded only marginally acceptable R^2 values (e.g., 0.806 with the FRM closest to it).

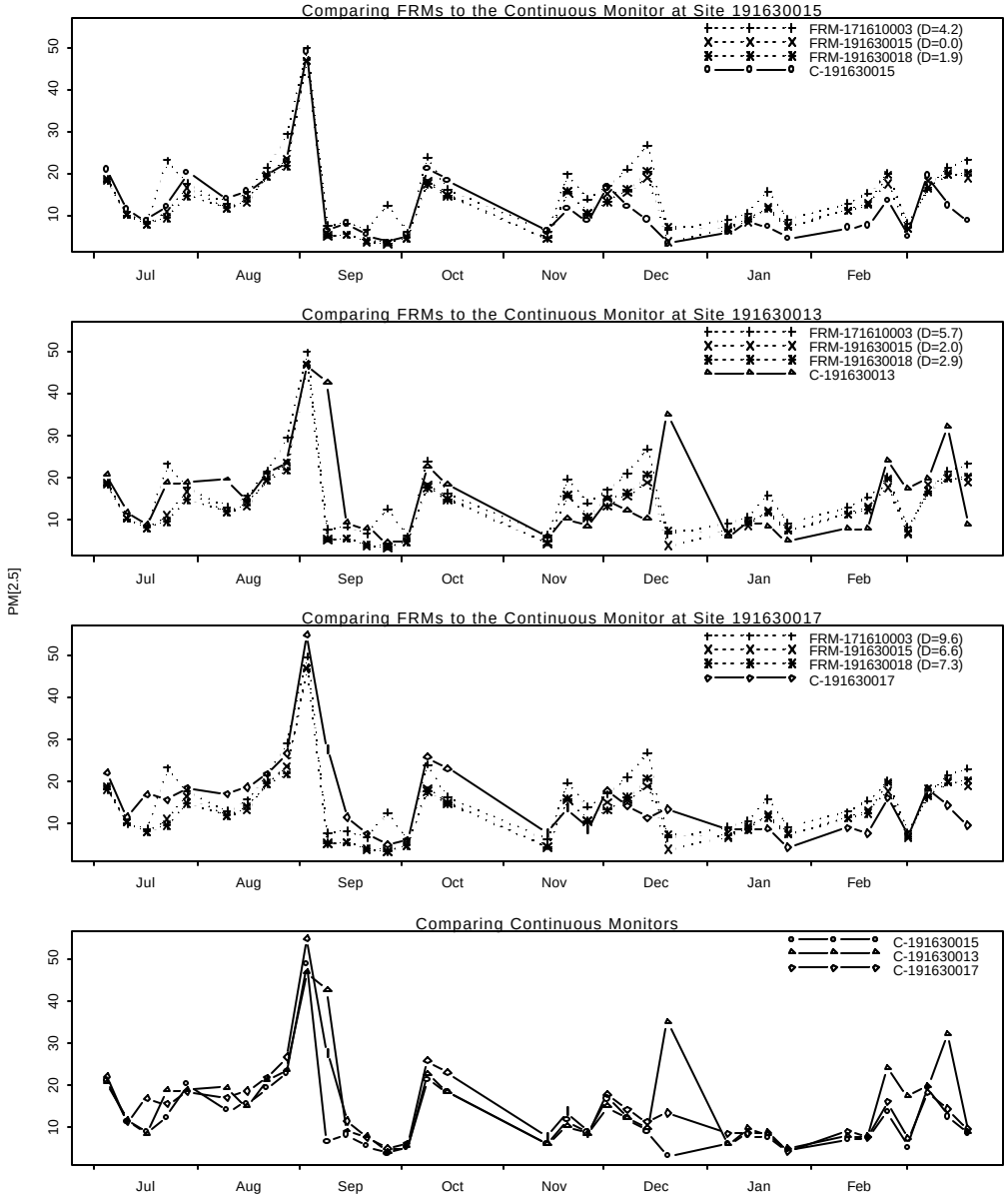

Figure B-15. **The three FRMs time series are compared to each $PM_{2.5}$ time series from a CM (the top three plots), and the three time series from the CMs are compared (bottom plot). The legend for the top three plots also shows the distance (D) from each of the FRM sites to the site with the CM in question.**

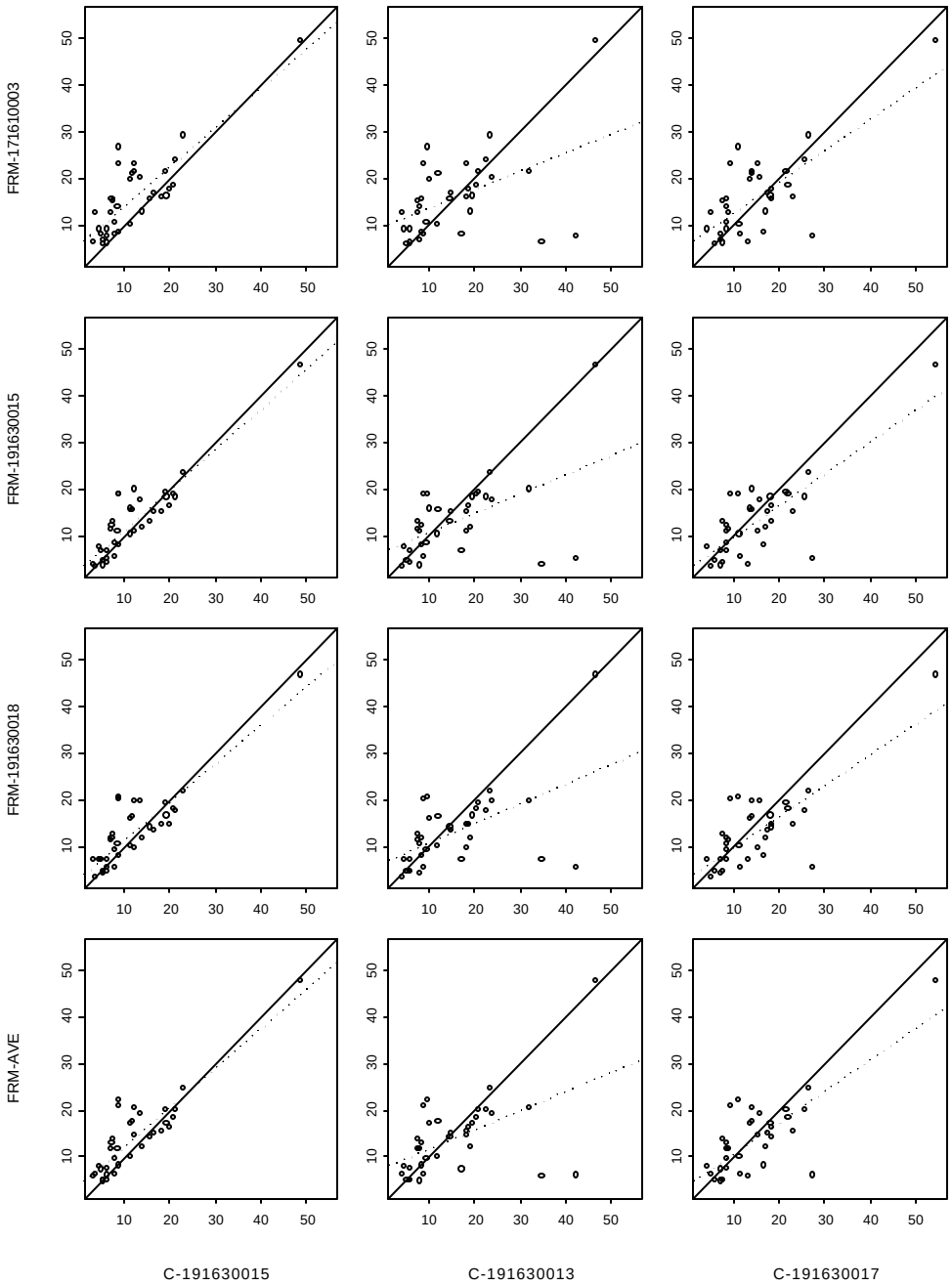

Figure B-16. Scatter plot of each FRM versus each CM. Also shown are the 45-degree line (solid) and a least squares model fit (dotted).

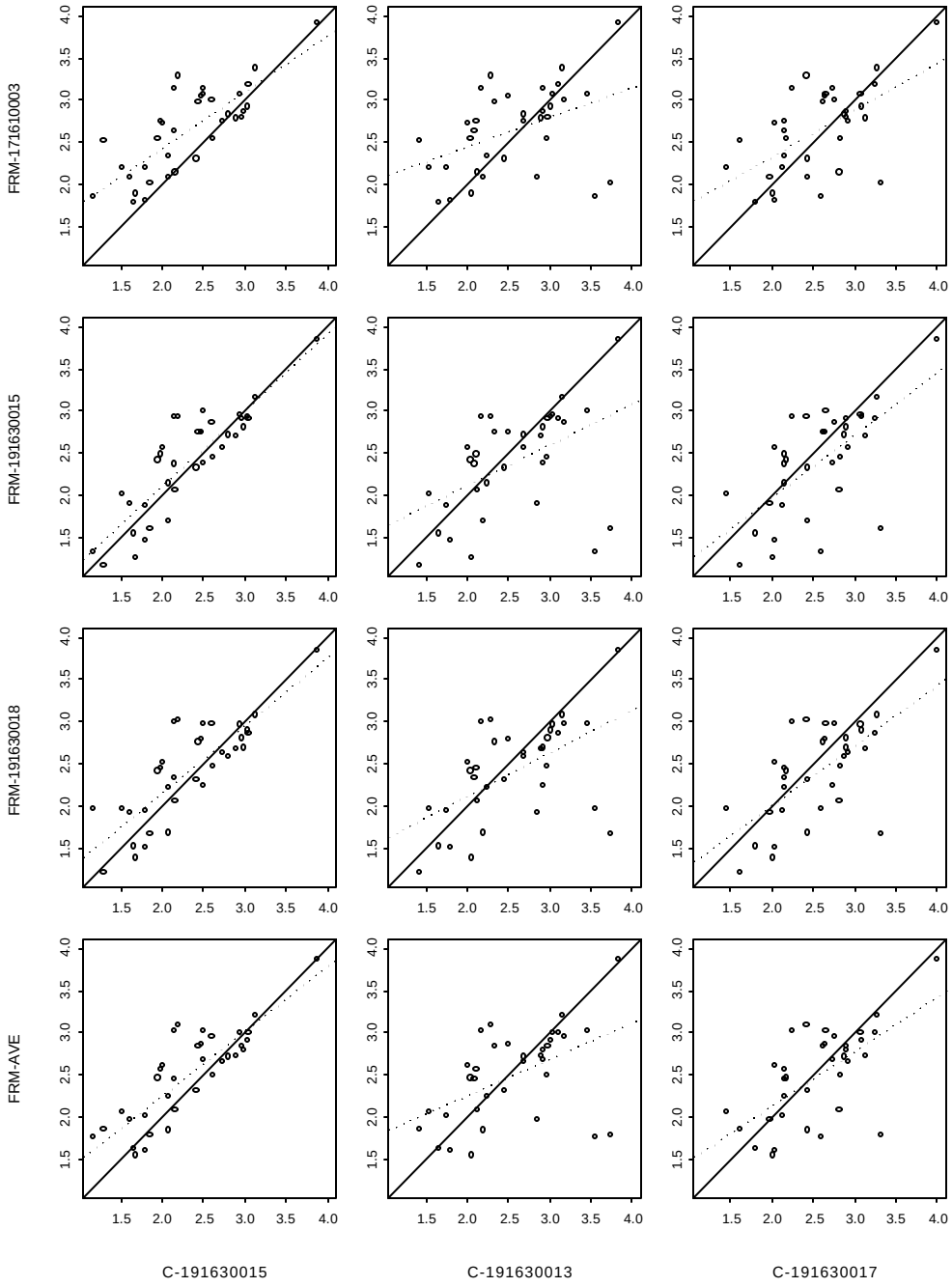

Figure B-17. Same as Figure B-16, except the $PM_{2.5}$ estimates have been log-transformed

Table B-4. Regression summary for a simple regression of each FRM (and their average) versus each CM in the Iowa-Illinois MSA.

CM	FRM	n	Interc.	se(Int.)	Slope	se(Sl.)	RMSE	R^2	Distance
191630013	191630015	35	6.527	2.173	0.413	0.115	7.019	0.282	1.957
	191630018	35	6.416	2.103	0.421	0.111	6.791	0.303	2.931
	171610003	35	9.682	2.383	0.395	0.126	7.696	0.230	5.676
	AVE	35	7.542	2.181	0.410	0.115	7.045	0.277	3.521
191630015	191630015	35	2.443	1.044	0.869	0.070	3.492	0.822	0.000
	191630018	35	2.989	1.168	0.825	0.079	3.909	0.769	1.900
	171610003	35	5.602	1.450	0.845	0.098	4.853	0.694	4.185
	AVE	35	3.678	1.162	0.846	0.078	3.887	0.780	2.028
191630017	191630015	35	2.780	1.652	0.684	0.094	5.116	0.618	6.560
	191630018	35	3.117	1.659	0.663	0.094	5.136	0.601	7.279
	171610003	35	5.795	1.929	0.674	0.109	5.974	0.536	9.626
	AVE	35	3.897	1.694	0.674	0.096	5.244	0.599	7.822

Table B-5. Same as Table B-4, except $PM_{2.5}$ estimates have been log-transformed.

CM	FRM	n	Interc.	se(Int.)	Slope	se(Sl.)	RMSE	R^2	Distance
191630013	191630015	35	1.153	0.397	0.483	0.150	0.550	0.238	1.957
	191630018	35	1.100	0.354	0.511	0.134	0.491	0.306	2.931
	171610003	35	1.743	0.336	0.350	0.127	0.465	0.187	5.676
	AVE	35	1.389	0.346	0.432	0.131	0.479	0.248	3.521
191630015	191630015	35	0.297	0.217	0.902	0.091	0.315	0.751	0.000
	191630018	35	0.551	0.230	0.801	0.096	0.334	0.678	1.900
	171610003	35	1.099	0.223	0.664	0.093	0.324	0.607	4.185
	AVE	35	0.720	0.208	0.765	0.087	0.302	0.702	2.028
191630017	191630015	35	0.509	0.397	0.736	0.151	0.482	0.417	6.560
	191630018	35	0.597	0.362	0.709	0.138	0.440	0.443	7.279
	171610003	35	1.234	0.344	0.550	0.131	0.417	0.347	9.626
	AVE	35	0.850	0.348	0.643	0.133	0.423	0.414	7.822

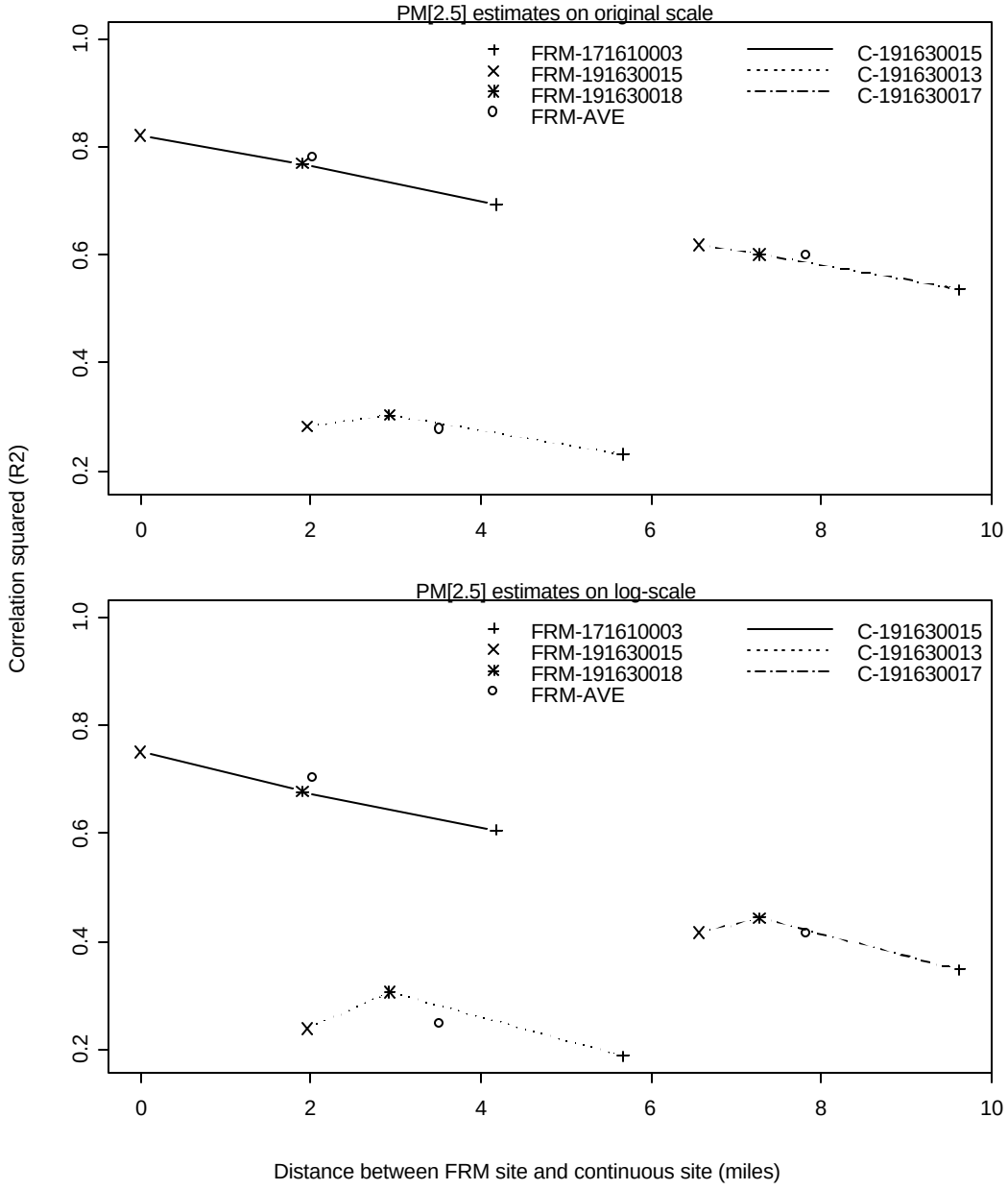

Figure B-18. R^2 between different FRMs (different symbols) and CMs (different line types) plotted versus the distance between the monitors. The two graphs correspond to untransformed $PM_{2.5}$ estimates (top) and log-transformed (bottom).

Conclusions

At the site with co-located continuous and FRM data (site 191630015), the difference between the FRM and the CM measurements showed seasonal patterns. After adjusting for the seasonal pattern, either by using a periodic seasonal trend or including temperature data, a satisfactory R^2 of 0.84 or better was achieved. The two continuous monitors not co-located with FRMs did not show strong enough correlation to nearby FRMs, even after adjusting for seasonality and removing possible outliers. The continuous monitor at site 191630013, which is only about two miles away from the co-located site, appeared problematic (see Tables B-4 and B-5 and Figure B-18).

B.3 SALT LAKE CITY-OGDEN, UTAH

Data from thirteen FRMs and two CMs are available from the Utah MSA, and the two CMs are co-located with FRMs. The data from the FRMs are from the beginning of 1999 through March 2000, but the data from the CMs are from the beginning of December 1999 through July 2000. There are only about four months of overlapping data, but the two co-located FRMs were sampled daily, resulting in a reasonable amount of data for analysis at the co-located sites. Figure B-19 shows the locations of the monitors.

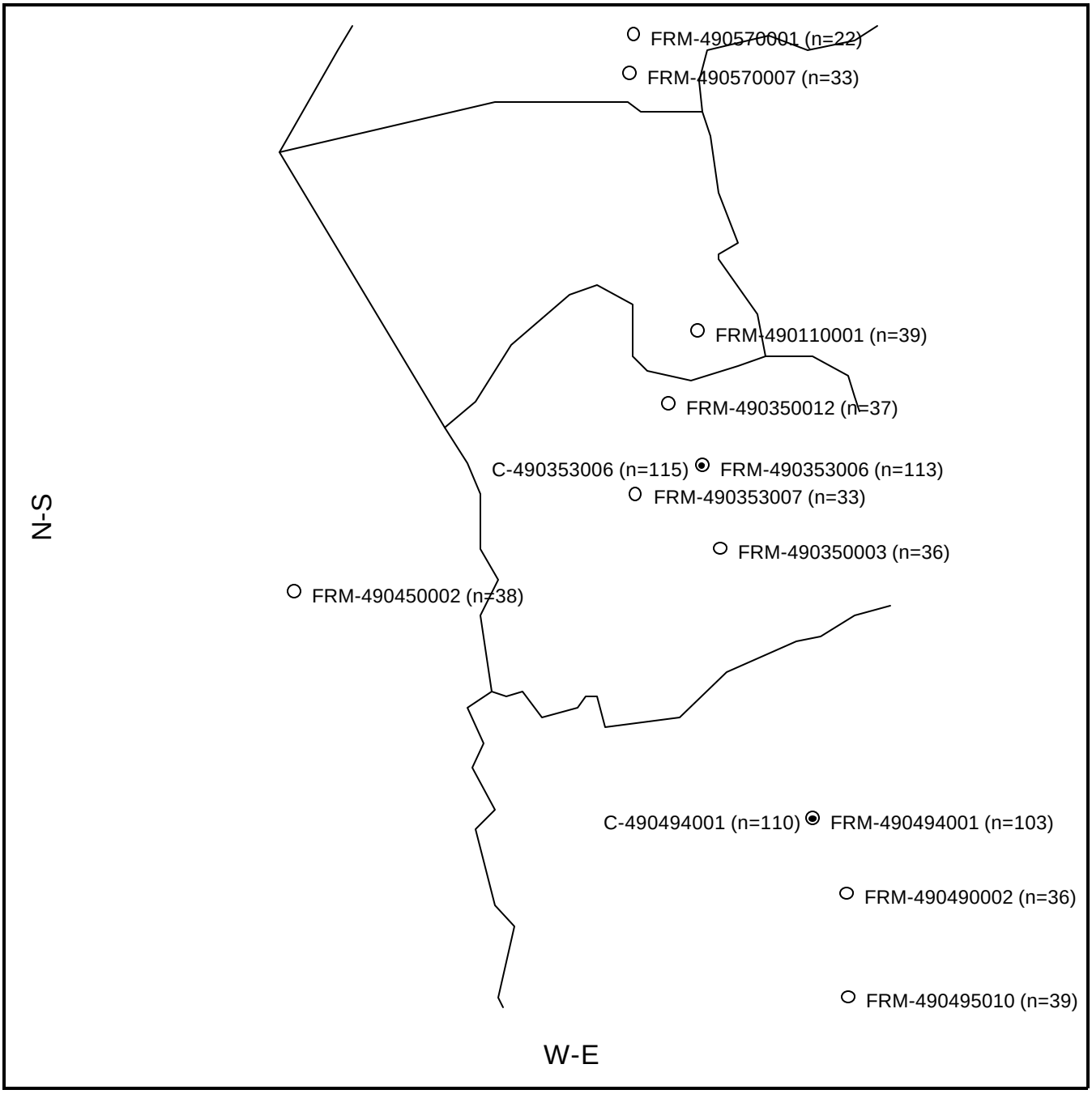

Figure B-19. Location of FRMs (circles) and CMs (black dots) sites. The number shown in parentheses is the number of PM2.5 observation available in the time period 12/01/1999 to 03/31/2000.

Analysis of Co-located Sites

The 490494001 co-located site has 97 days with observations from both a continuous and FRM monitor, covering the time period from 12/01/1999 to 03/31/2000. Over this same period, the 490253005 site has 111 observations. Figure B-20 shows the $PM_{2.5}$ time series and the scatter plots are given in Figure B-21. The time series plots show how the CMs underestimate the FRMs; but, on the other hand, the scatter plots show that this underestimation is systematic (i.e., consistent) and can therefore be corrected through least squares regression calibration. When log-transformed, the relationship between the FRM and the CM observations seems to be non-linear and three observations deviate from the main body of observations at both sites. Therefore, unlike what was observed for the North Carolina and Iowa-Illinois MSA's, a natural log-transformation may not be appropriate in the case of Utah. A summary of least square regression fits is given in Table B-6.

Table B-6. Summary of least squares regressions when regressing FRM versus CM measurements at the co-located sites in Utah.

		N	Intercept	se(Int.)	Slope	se(Slope)	R^2
untransformed	490494001	97	-3.233	0.972	1.485	0.074	0.808
	490353006	111	-3.368	0.854	1.612	0.063	0.858
log-transformed	490494001	97	0.175	0.214	0.942	0.092	0.526
	490353006	111	-0.289	0.182	1.173	0.079	0.669

When using untransformed data, both sites yield R^2 values above 0.8. The main reason for lower R^2 values when using log-transformed data is due to three outliers in both cases (see Figure B-21, scatter plots, and Figure B-22, histogram of residuals). The histogram of the residuals from the least squares regression model fit does not show strong evidence of skewness (Figure B-22). Since the period in question covers only four months, it is very difficult to check for seasonal changes in the relationship between the FRMs and the CMs. Figure B-23 shows the residuals from the least squares regressions plotted versus time, and there is some evidence of a decreasing trend over the four-month period.

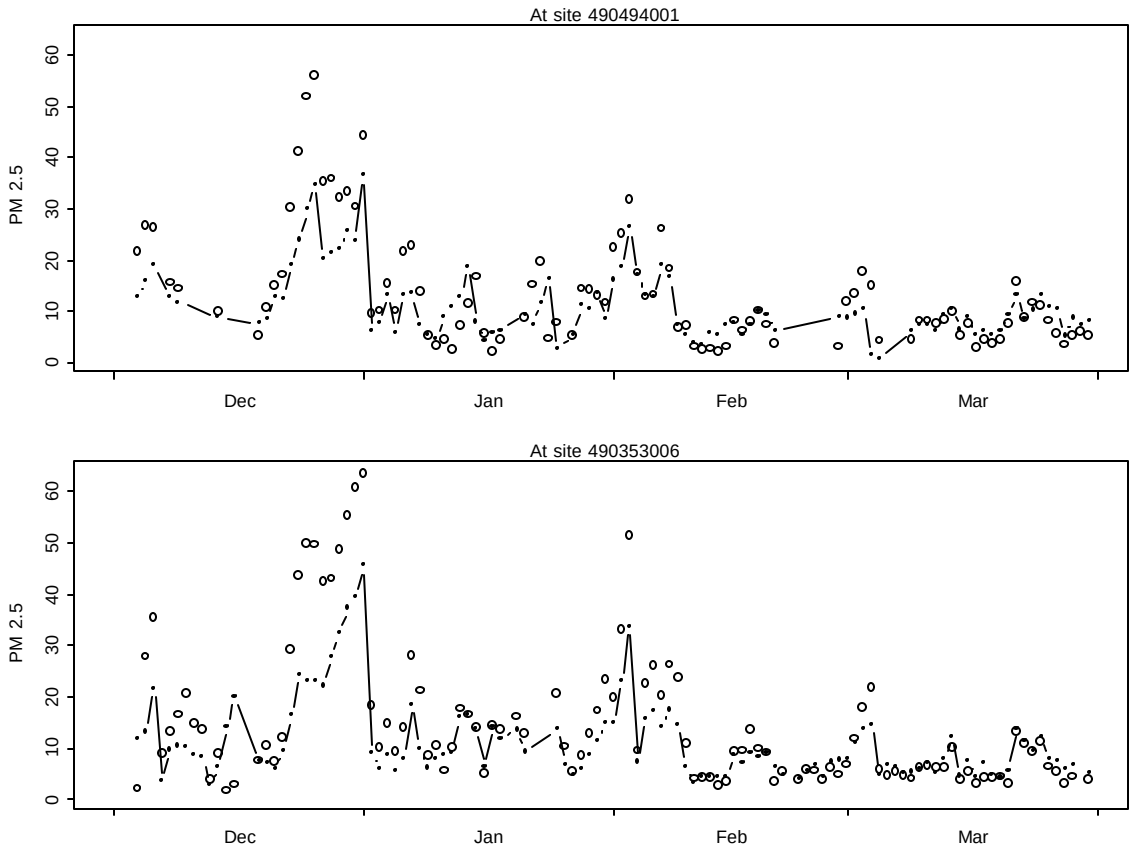

Figure B-20. Time series of $PM_{2.5}$ concentrations at the co-located Utah sites. The FRMs are circles and the CMs are dots connected with line.

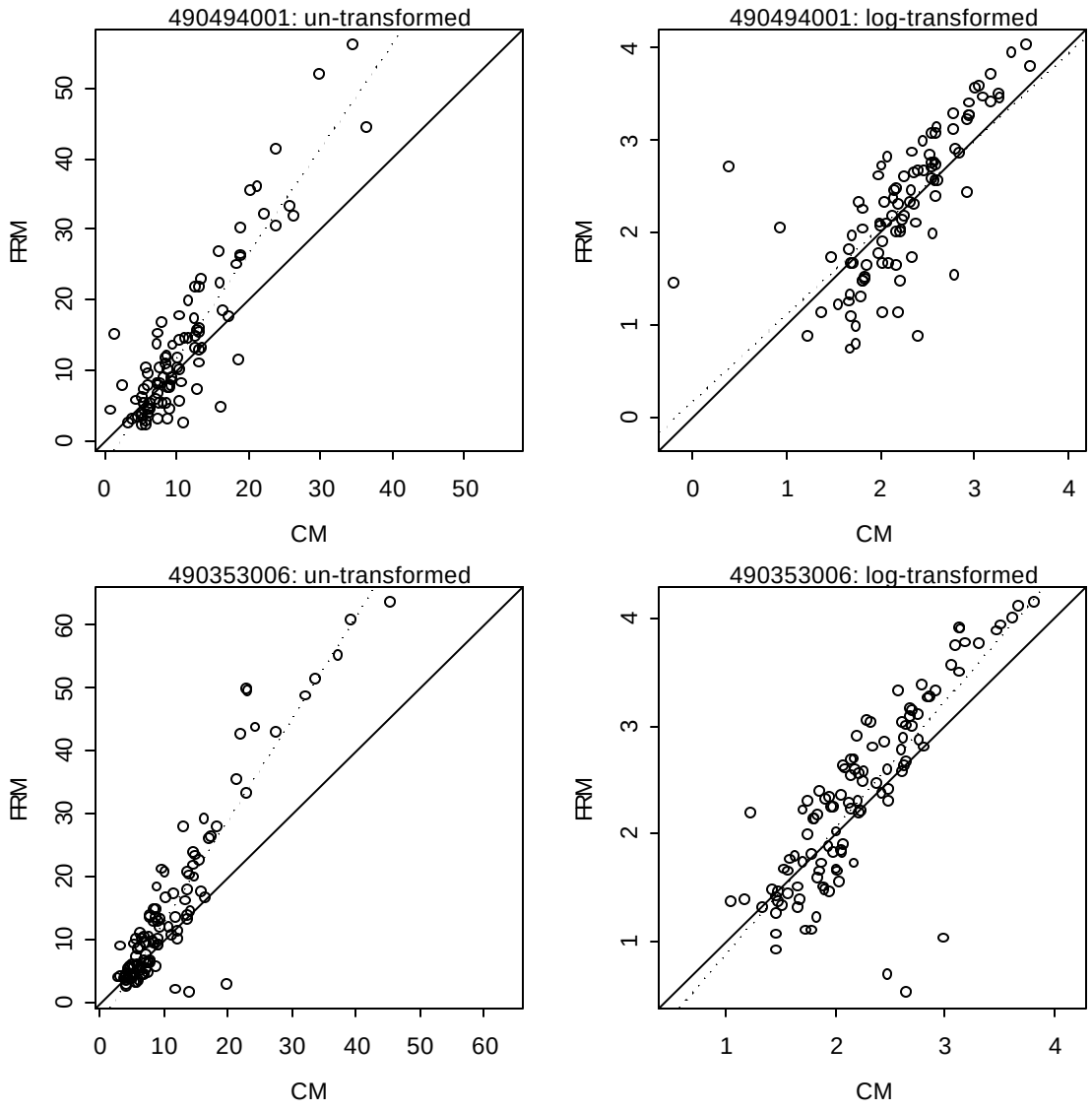

Figure B-21. Scatter plots of FRM values versus CM values at the two co-located Utah sites, for untransformed and log-transformed $PM_{2.5}$ concentrations. The solid line shows the 45-degree line and the dotted line is the least squares regression line.

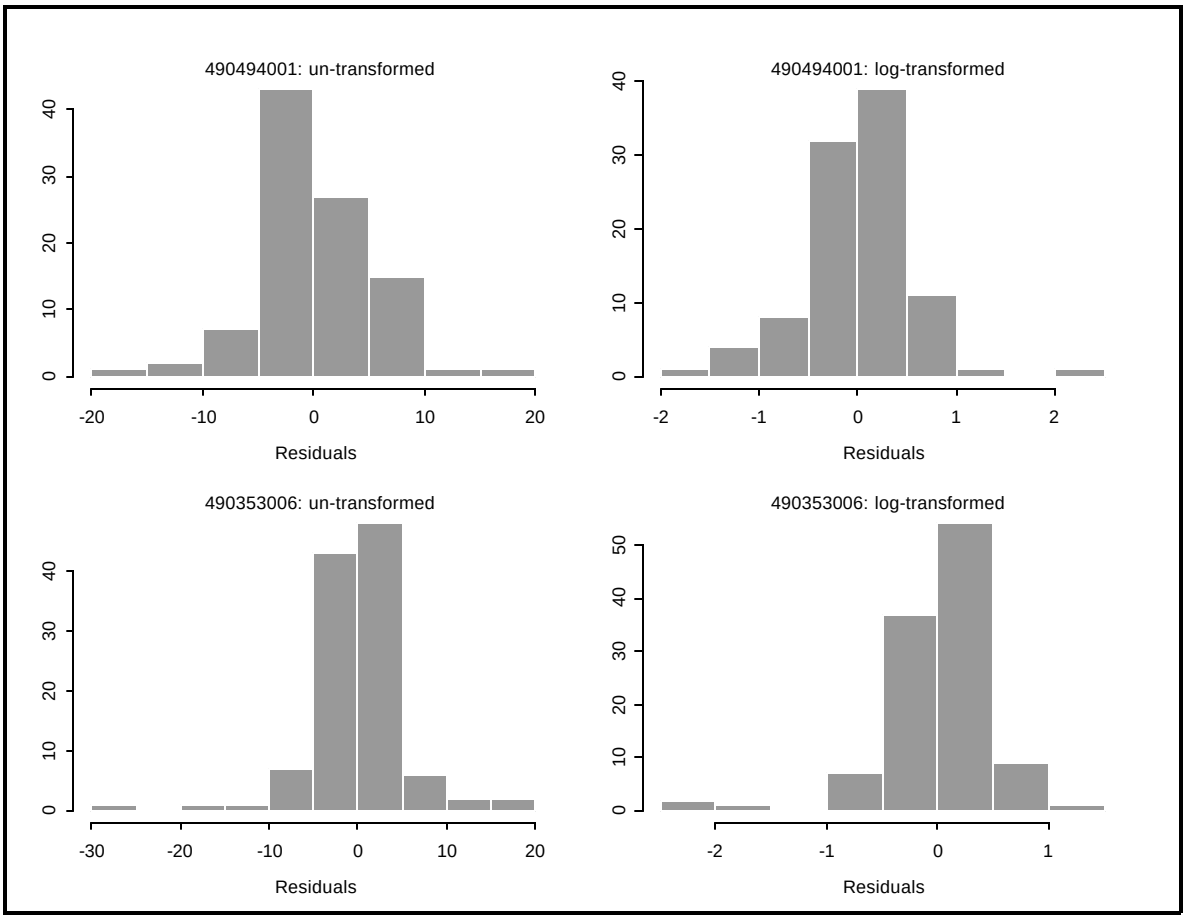

Figure B-22. Histogram of the residuals from the least squares regressions of FRM verus CM measurements at the two co-located Utah sites, for both untransformed and log-transformed $PM_{2.5}$ concentrations

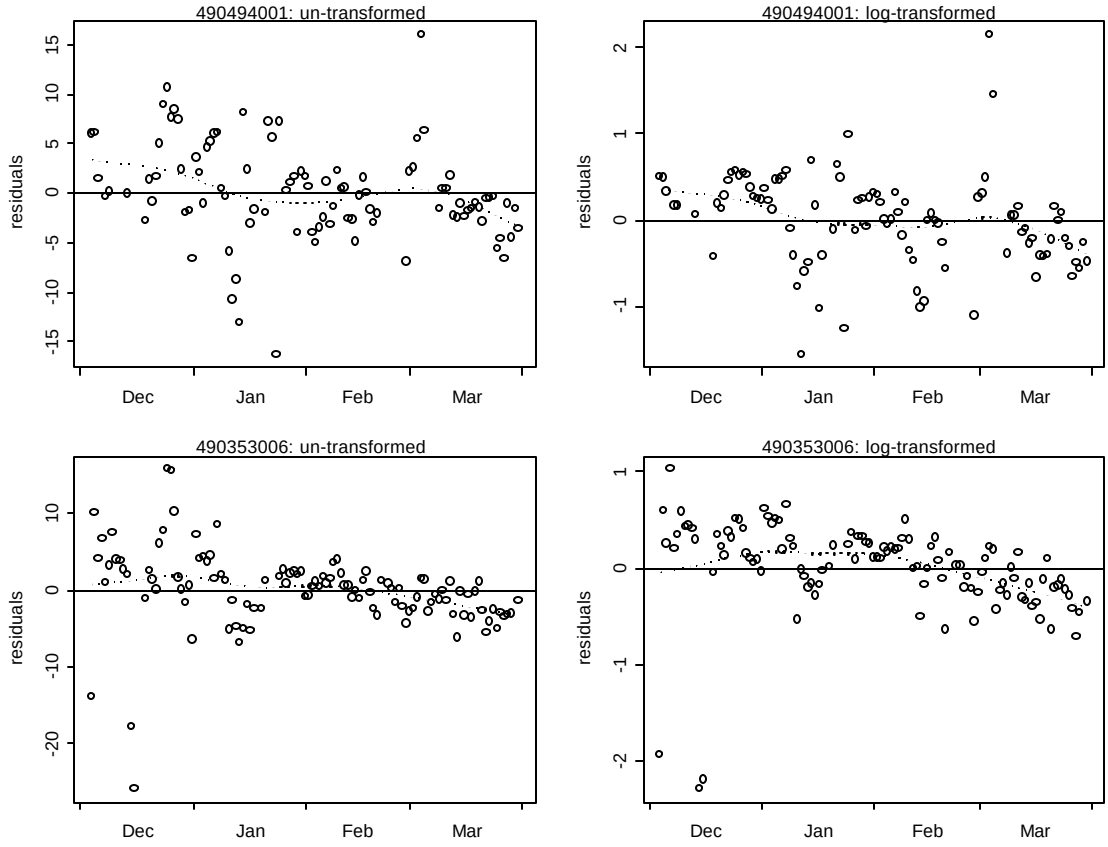

Figure B-23. Residuals from least squares regressions of FRM versus CM data at the two co-located Utah sites plotted versus time, for both untransformed and log-transformed data. The dotted line shows a smooth trend.

Analysis of Other Available Data

All 13 FRMs were compared to the two CMs by using all data available in the time period 12/01/1999 to 03/31/2000 for each FRM-CM pair. In general, only the sites with co-located continuous and FRM data possessed a large sample size of days for comparison. The largest sample size for a continuous-FRM not co-located was 39. Given the relatively short time period of observed data (approximately four months) and the relatively small sample sizes of available data ($n < 40$), model development based on non co-located continuous-FRM data is not highly recommended in this case. However, the following generalities were observed in the data:

- The time series and the scatter plots (Figures B-24 through B-27) show the same general underestimation pattern as was seen for the co-located sites.

- Figure B-28 shows R^2 plotted versus distance between each FRM-CM pair, and demonstrates a reasonably strong correlation in general for sites up to 20 miles away.

Conclusions

Because of the relatively short time period of observed data (approximately four months) and the relatively small sample sizes of available data ($n < 40$), model development for the Utah MSA probably should only be pursued for the two sites with co-located continuous and FRM data. In both cases, the regression models for the untransformed data appear more appropriate than those for the natural log-transformed data. Adjustments for seasonality are extremely limited given the short time period over which the data are observed. However, the basic models for the untransformed data at the two co-located Utah sites, which do not adjust for seasonality or meteorological data, yield R^2 values above 0.8. Given that around 100 observations were used to develop these models, Tables 2-2 and 2-3 of Chapter 2 suggest they may be reasonable for use along with continuous $PM_{2.5}$ measurements to report an AQI in the Utah MSA.

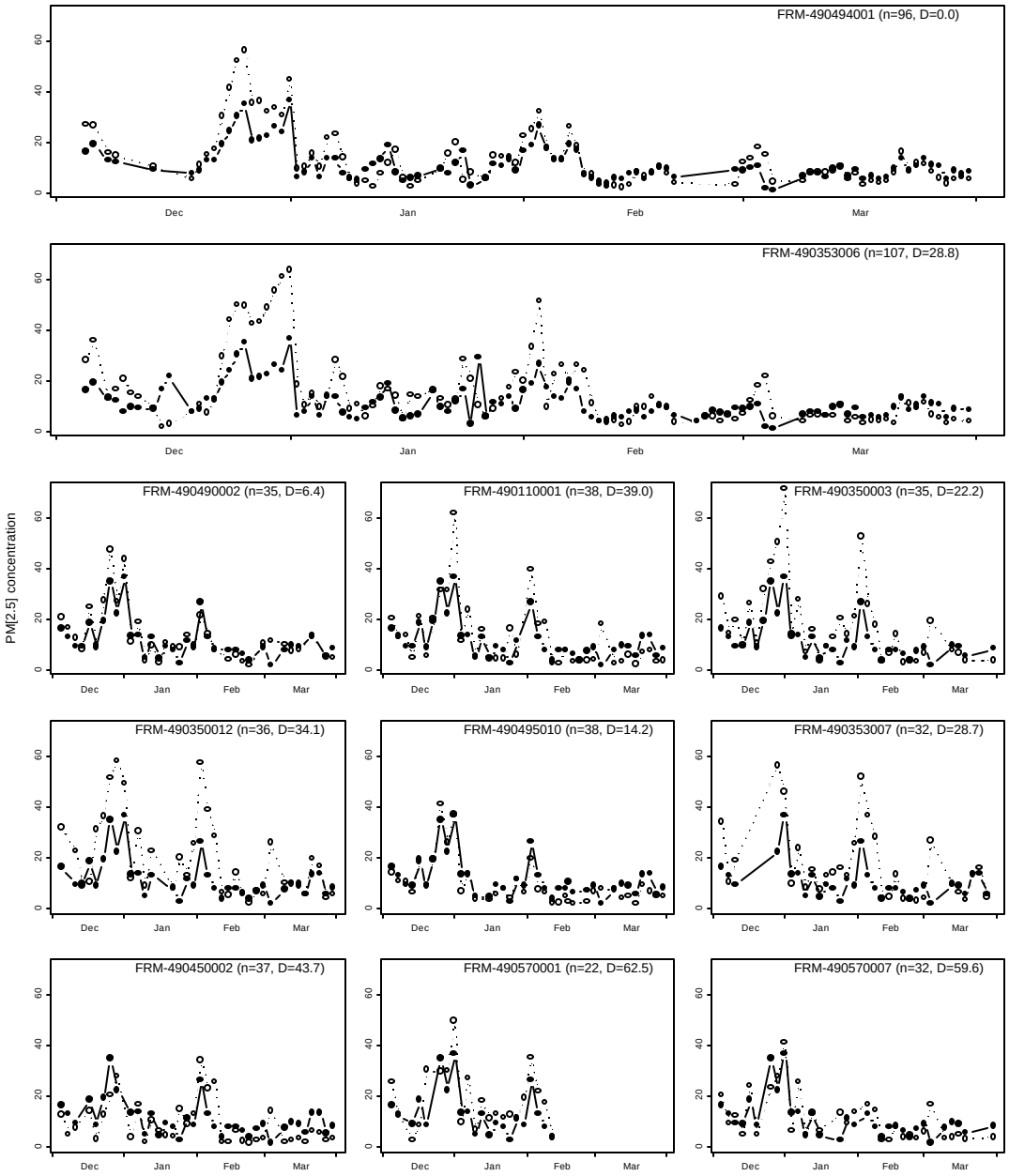

Figure B-24. **Each panel shows one FRM $PM_{2.5}$ time series (circles) and the time series from the CM at site 49049001 (black dots). Each panel is labeled with the FRM in question, the number of observations (n) and the distance between the FRM and the continuous monitor (D).**

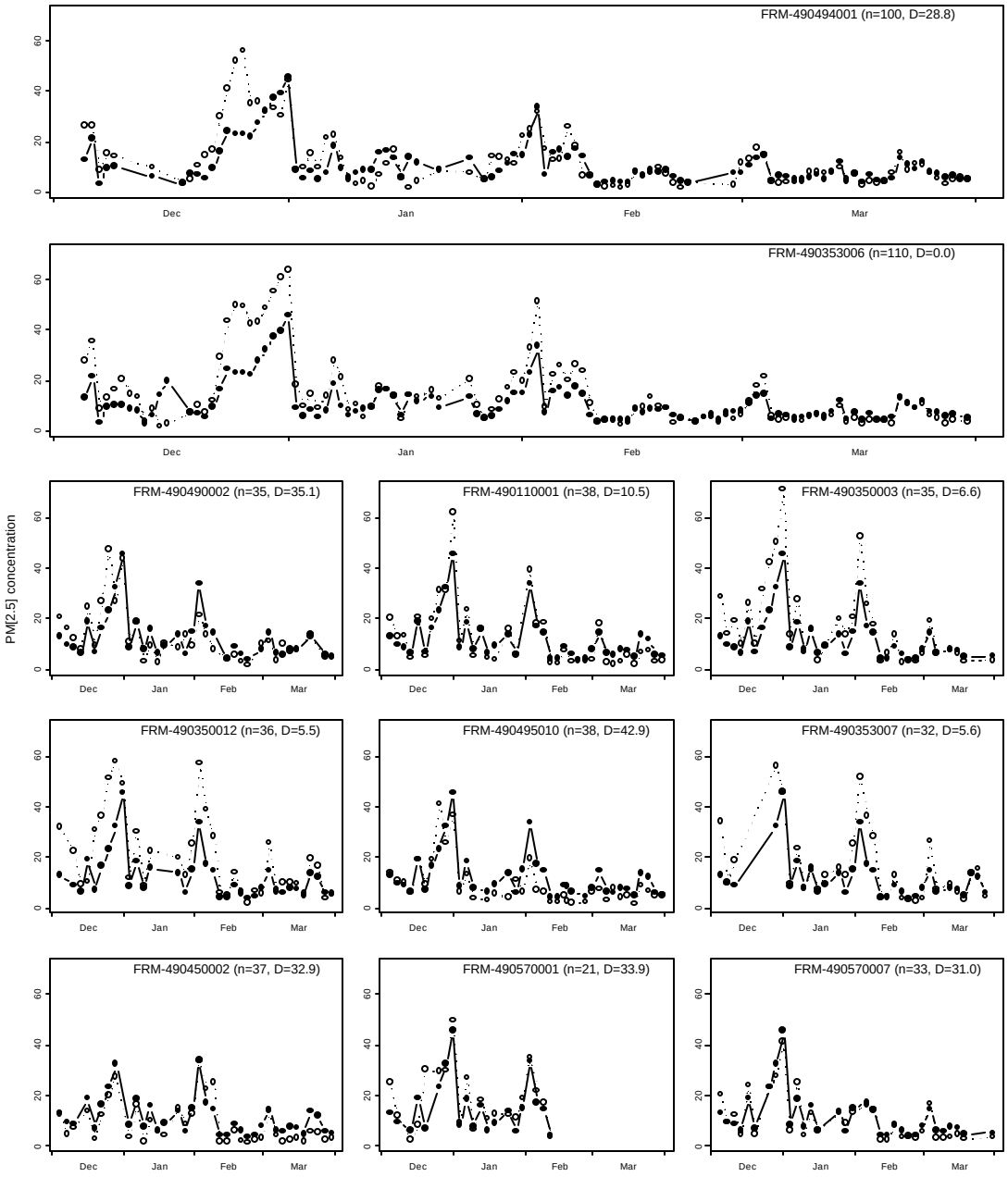

Figure B-25. Identical to Figure B-24, but for the CM at site 490353006 (black dots).

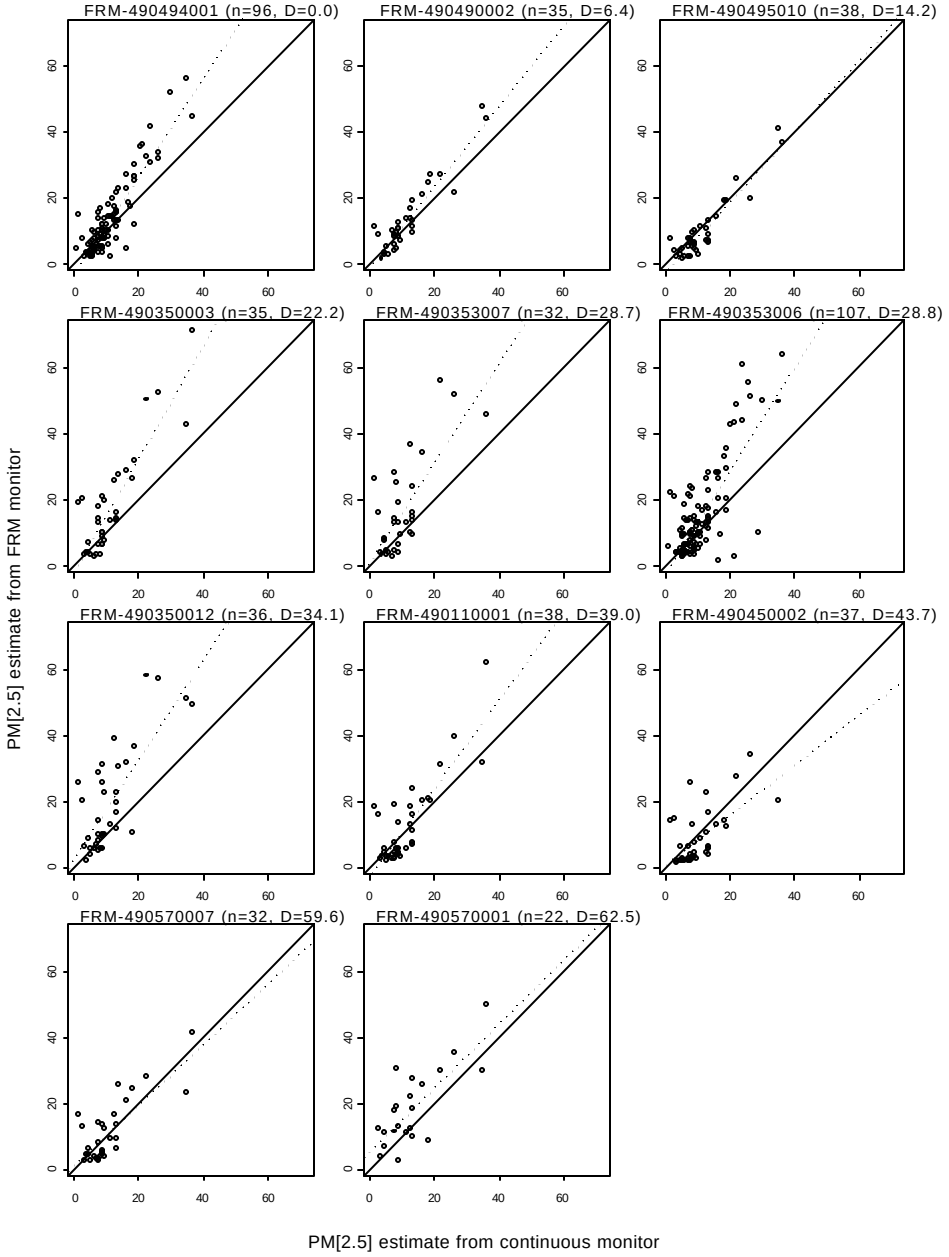

Figure B-26. Each panel shows a scatter plot of $PM_{2.5}$ estimates from an FRM monitor versus the estimates derived from the CM at site 49049001, along with the 45-degree line (solid) and a least squares regression fit (dotted). Each panel is labeled with the FRM site in question, the number of observations (n), and the distance between the FRM and the CM (D).

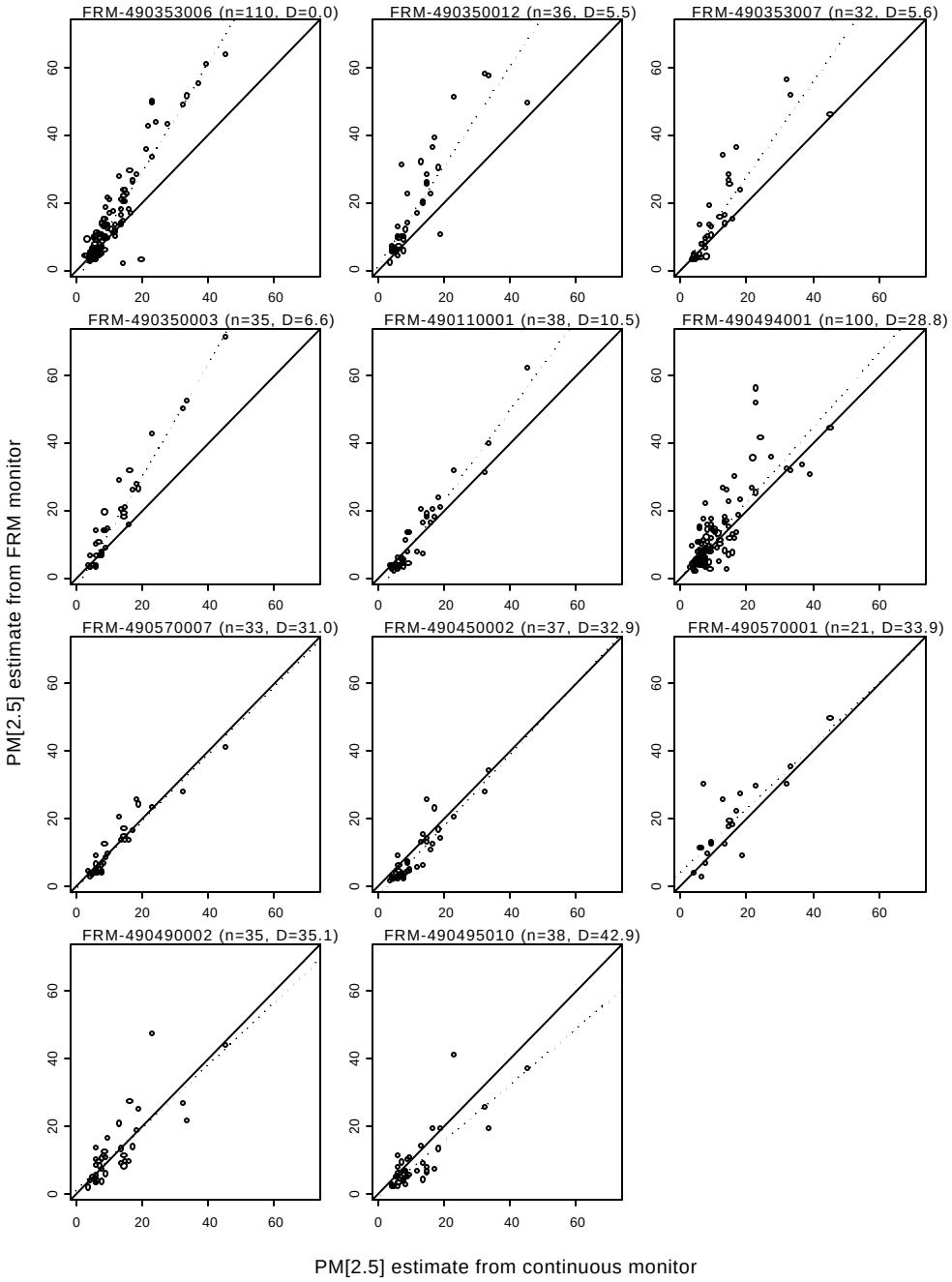

Figure B-27. Identical to Figure B-26, but the for the CM at site 490353006.

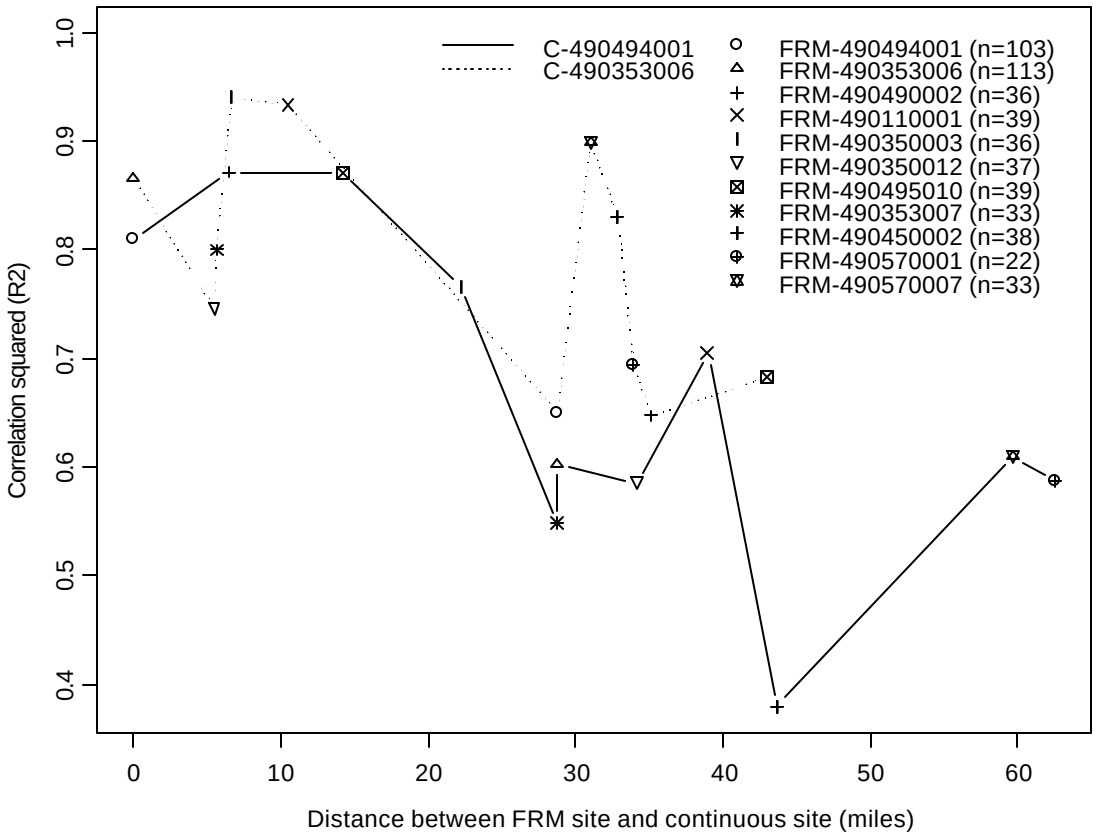

Figure B-28. **R^2 between various FRMs (different symbols) and CMs (different lines) plotted versus the distance between the two monitors (untransformed data)**

B.4 HOUSTON, TEXAS

There are eight FRM monitors and four continuous monitors (CMs) in the Houston, Texas MSA. Three sites have co-located FRMs and CMs. One FRM site has only nineteen observations and is not considered in any of the analyses performed. Henceforth, we only consider seven FRMs and four CMs.

Figure B-29 shows the location of the monitors along with the total number of observations in the period from 02/01/00 to 06/30/00. Only six days have observations from all eleven monitors. When only looking at the three co-located sites, only 12 days have observations from all five monitors. The approach taken therefore is to start with a small study of the three co-located sites (ignoring all spatial relationships) and follow with a more in-depth study comparing three FRMs to two CMs using days where all five monitors have observations.

Analysis of Co-located Sites

Figure B-30 shows the time series of $PM_{2.5}$ estimates from both the FRMs and the CMs at the three co-located sites, and Figure B-31 shows the scatter plots. Figure B-31 clearly shows a systematic bias in the CMs (as well as some outliers). The 45-degree solid lines and least squares regression dashed lines in Figure B-31 are parallel but vertically shifted apart from one another. The bias is confirmed in the least squares regression summaries of Tables B-7 and B-8. While the slopes are very near one on the untransformed scale, the intercepts are all clearly above zero.

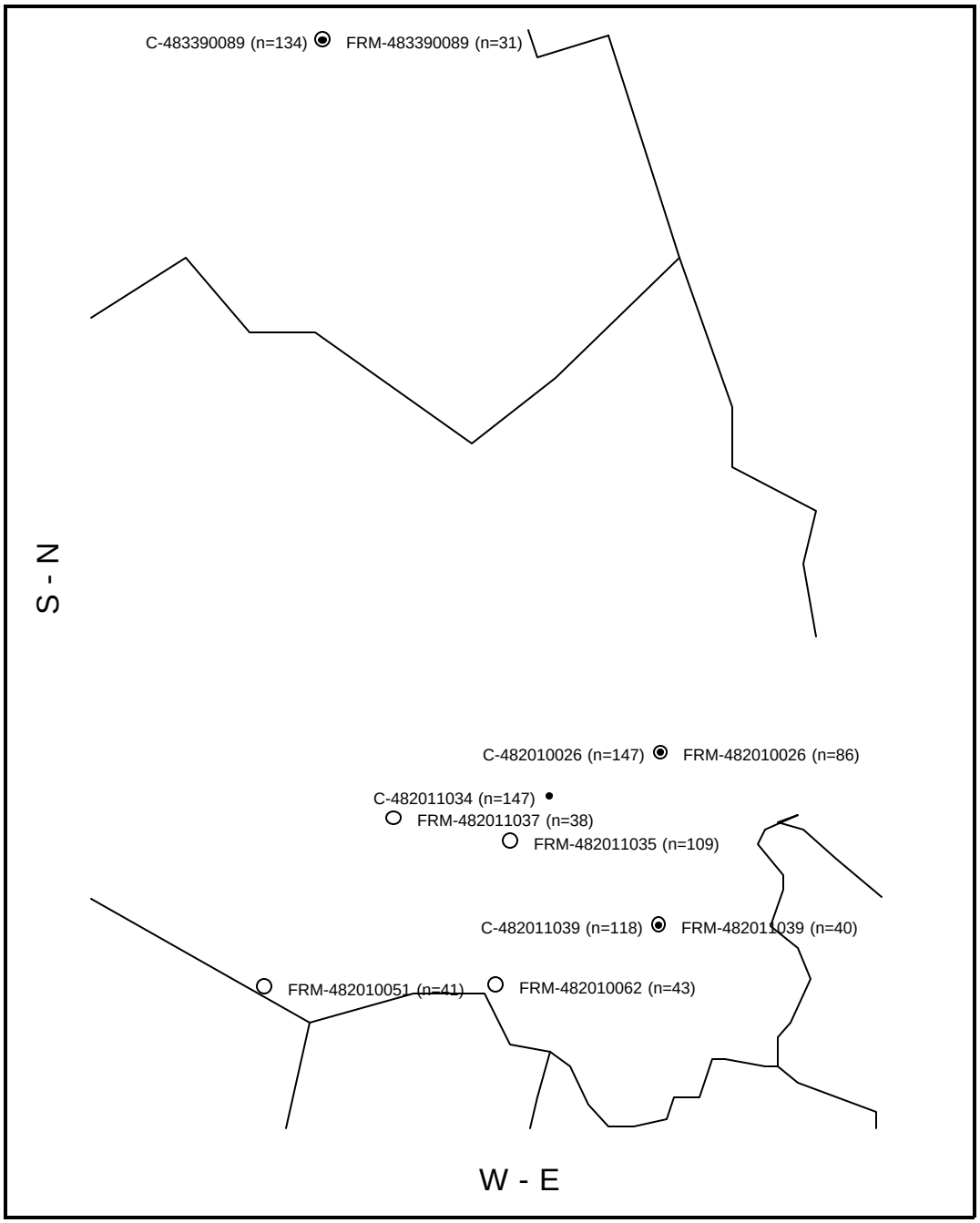

Figure B-29. Location of the seven FRMs and four CMs in the Texas MSA. The number in parentheses shows the number of observations available from 02/01/00 to 06/30/00.

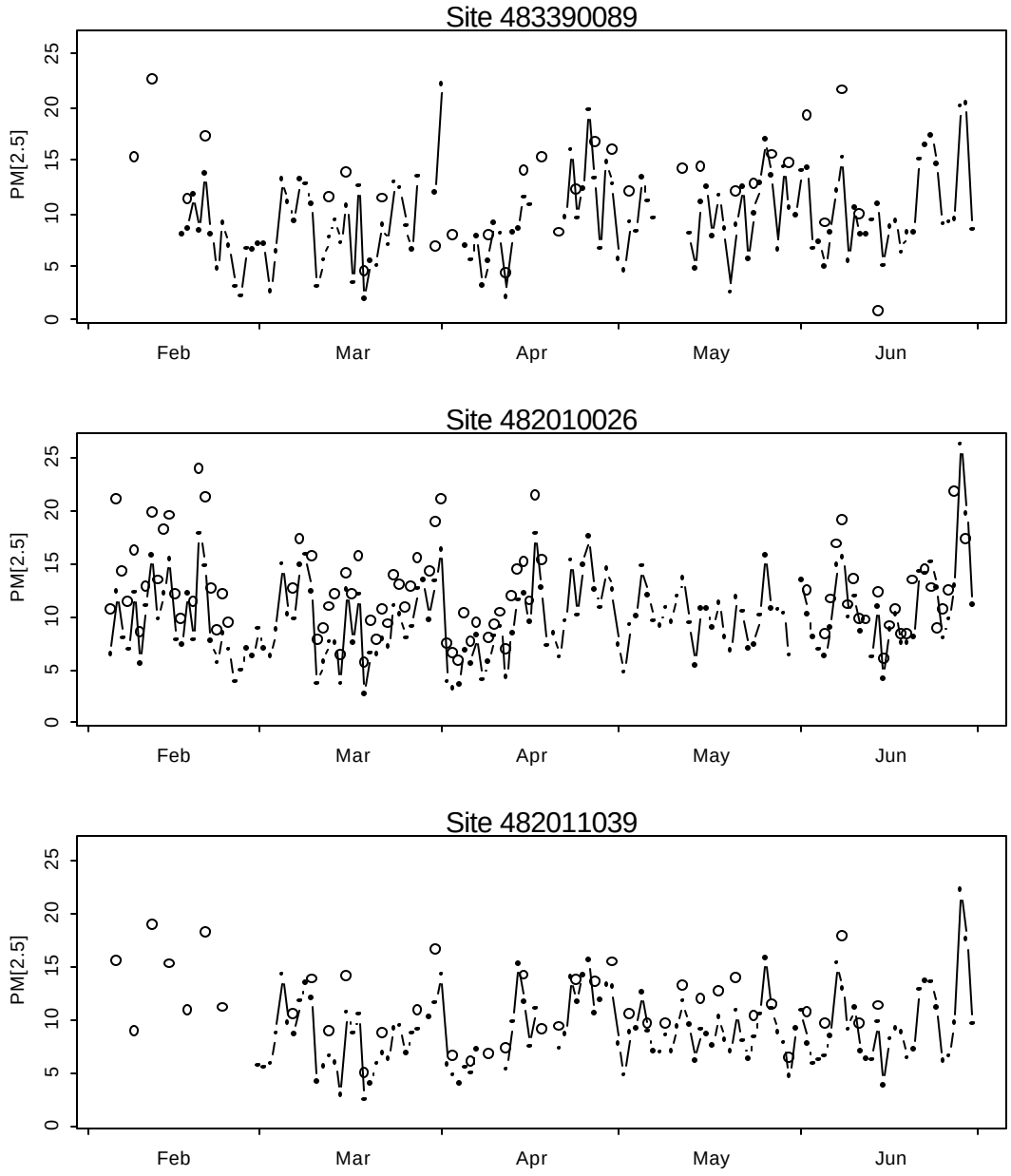

Figure B-30. Time series of $PM_{2.5}$ values at the three co-located Texas MSA sites. FRM values are displayed as circles and the CM values as black dots connected with a solid line if observed on consecutive days.

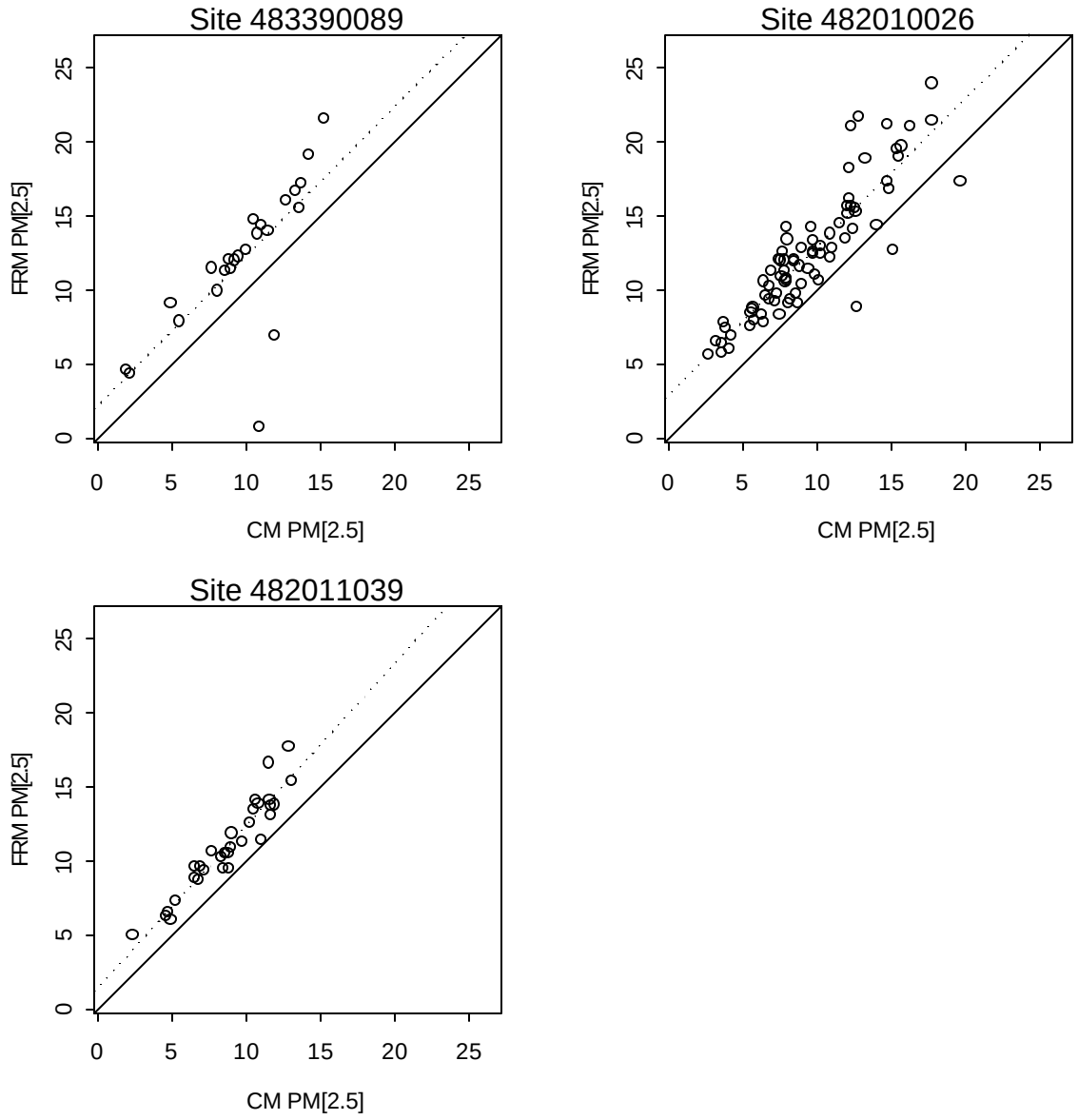

Figure B-31. **Scatter plot of FRM $PM_{2.5}$ values versus CM $PM_{2.5}$ values at the three co-located sites. The solid line shown is the 45-degree line and the dashed line is a simple least squares regression line.**

Table B-7. Least squares regression summaries based on the three Texas sites with co-located continuous and FRM data, on the original untransformed $PM_{2.5}$ scale.

Site	N	Intercept	se(Int.)	Slope	se(Slope)	R^2	RMSE
483390089	24	2.202	2.059	1.008	0.198	0.540	3.347
482010026	82	2.925	0.608	0.999	0.059	0.780	1.993
482011039	31	1.432	0.619	1.093	0.067	0.901	0.997

Table B-8. Least squares regression summaries based on the three Texas sites with co-located continuous and FRM data, on the log-transformed $PM_{2.5}$ scale.

Site	N	Intercept	se(Int.)	Slope	se(Slope)	R^2	RMSE
483390089	24	1.086	0.544	0.580	0.243	0.206	0.618
482010026	82	0.915	0.086	0.714	0.039	0.811	0.147
482011039	31	0.724	0.098	0.773	0.046	0.908	0.094

The 483390089 site demonstrates a low correlation between continuous and FRM data compared to the other two sites, which can easily be explained by the two outliers seen in Figure B-31. Removing the two outliers in question increases the R^2 for the site's model to above 0.9. However, the site contains few days of observations to begin with (n=24) and a clear justification for removing the two apparent outliers was not available. The results for the other two sites (482010026 and 482011039) are encouraging. R^2 values are slightly higher for these two sites when using log-transformed data to develop their associated model. Using log-transformed data, the R^2 value of both of these sites is above 0.8, which is achieved without conducting any further model development to adjust for seasonality or meteorological data.

Analysis of Other Available Data

Of the seven FRMs and four CMs, three FRMs and two CMs were identified as providing a reasonable number of days for which all monitors have $PM_{2.5}$ measurements. This resulted in 22 days

sampled by all five monitors in the time period 02/01/00 to 06/30/00. The chosen sites with FRM monitors are 482011035, 482010026, and 482010062. The chosen sites with continuous monitors are 482010026 and 482011034.

Figure B-32 shows the time series of the $PM_{2.5}$ estimates from these five monitors; comparing the three FRMs versus each CM, and then comparing the two CMs with one another. Figure B-33 shows the scatter plots (each FRM versus each CM). The scatter plots identify obvious outliers and bias (intercept different from zero and slope different from one). A least squares regression summary is given in Table B-9 for the original $PM_{2.5}$ concentration scale, and in Table B-10 when the data are log-transformed. Looking at Table B-10, R^2 values remain reasonably high (above 0.63) at distances up to 15 miles.

Figure B-34 more clearly shows the R^2 values plotted versus distance (based on the results summarized in Tables B-9 and B-10). Figure B-34 can be compared to Figure B-35, which shows the correlation between the two CMs (one number) and the correlation between the three FRMs (three comparisons). The continuous monitors appear to correlate quite well with one another, even though they're separated by a distance of about 6 miles. This may be suggestive of a relatively high level of precision associated with these two monitors. The FRM monitors do not fare as well with respect to their correlation with one another. However, these monitors are separated by even greater distances (approximately 8 miles or more).

Conclusions

Focusing on co-located data, the 483390089 site has several outliers and few data points (n=24). Therefore, model development at this site probably should not be pursued using the currently available data. Results for the other two co-located sites (482010026 and 482011039) are more encouraging. R^2 values are slightly higher for these two sites when using log-transformed data to develop their associated model. Using log-transformed data, the R^2 value of both of these sites is

above 0.8, which is achieved without conducting any further model development to adjust for seasonality or meteorological data.

Specifically, the model for log-transformed co-located continuous-FRM measurements at the 482011039 site has an R^2 value of 0.908, based on n=31 observations. According to Tables 2-2 and 2-3 of Chapter 2, this model is acceptable (depending on the Houston decision-makers' tolerable levels of decision errors) for use along with continuous $PM_{2.5}$ measurements to report AQI values in the Texas MSA. Finally, according to Table B-10, a similar acceptable model might be applied, which relates continuous $PM_{2.5}$ measurements to the average of several nearby FRM measurements (within 15 miles). This conclusion, however, is based on only n=22 observations.

Table B-9. Least squares regression summary of each of the three FRMs, and their average, versus each of the two CMs, on the original $PM_{2.5}$ concentration scale. The last column, Dist., is the distance between the monitors in miles.

CM	FRM	N	Interc.	se(Int.)	Slope	se(Sl.)	RMSE	R^2	Dist.
482010026	482010026	22	2.964	1.237	1.011	0.110	2.116	0.808	0.000
	482011035	22	5.945	1.754	0.832	0.156	3.000	0.587	9.241
	482010062	22	4.491	1.318	0.638	0.117	2.254	0.598	15.172
	AVE	22	4.466	1.108	0.827	0.099	1.895	0.779	8.138
482011034	482011035	22	5.076	1.762	0.865	0.149	2.853	0.627	3.155
	482010026	22	2.080	1.203	1.036	0.102	1.949	0.837	6.304

Table B-10. Least squares regression summary of each of the three FRMs, and their average, versus each of the two CMs, on the log-transformed scale. The last column, Dist., is the distance between the monitors in miles.

CM	FRM	N	Interc.	se(Int.)	Slope	se(Sl.)	RMSE	R^2	Dist.
482010026	482010026	22	0.829	0.132	0.759	0.057	0.125	0.898	0.000
	482011035	22	1.218	0.214	0.626	0.093	0.203	0.695	9.241
	482010062	22	1.179	0.207	0.527	0.090	0.196	0.632	15.172
	AVE	22	1.100	0.139	0.633	0.060	0.132	0.846	8.138
482011034	482011035	22	1.163	0.227	0.633	0.096	0.206	0.684	3.155
	482010026	22	0.770	0.152	0.765	0.064	0.138	0.876	6.304
	482010062	22	1.049	0.192	0.569	0.081	0.175	0.710	10.449

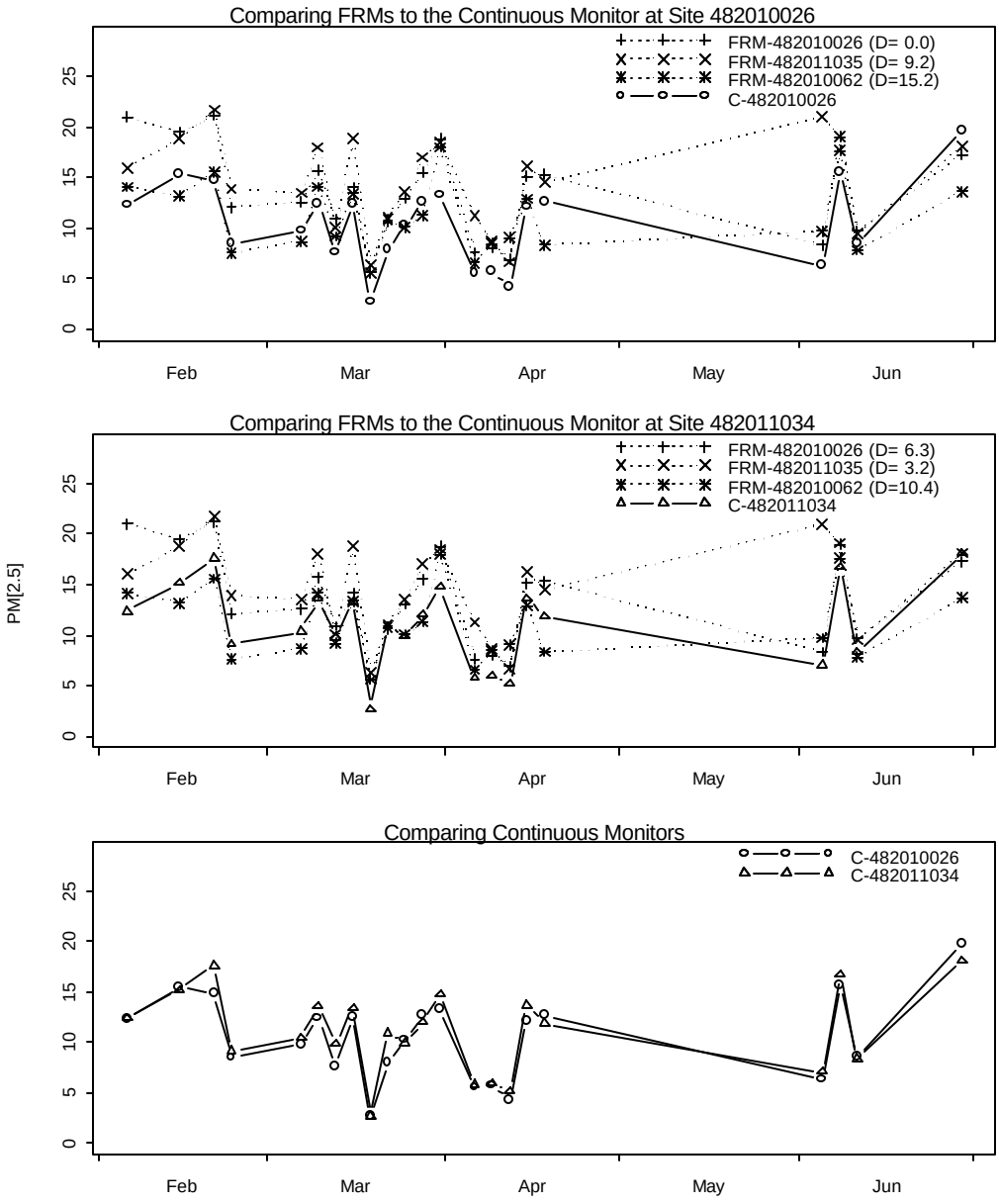

Figure B-32. **The three FRM $PM_{2.5}$ time series are compared to each $PM_{2.5}$ time-series from a CM (the top two plots) and the two $PM_{2.5}$ time series from the CMs are compared with one another (bottom plot). The legend for the top two plots also shows the distance (D) from the FRM sites to the CM site in question.**

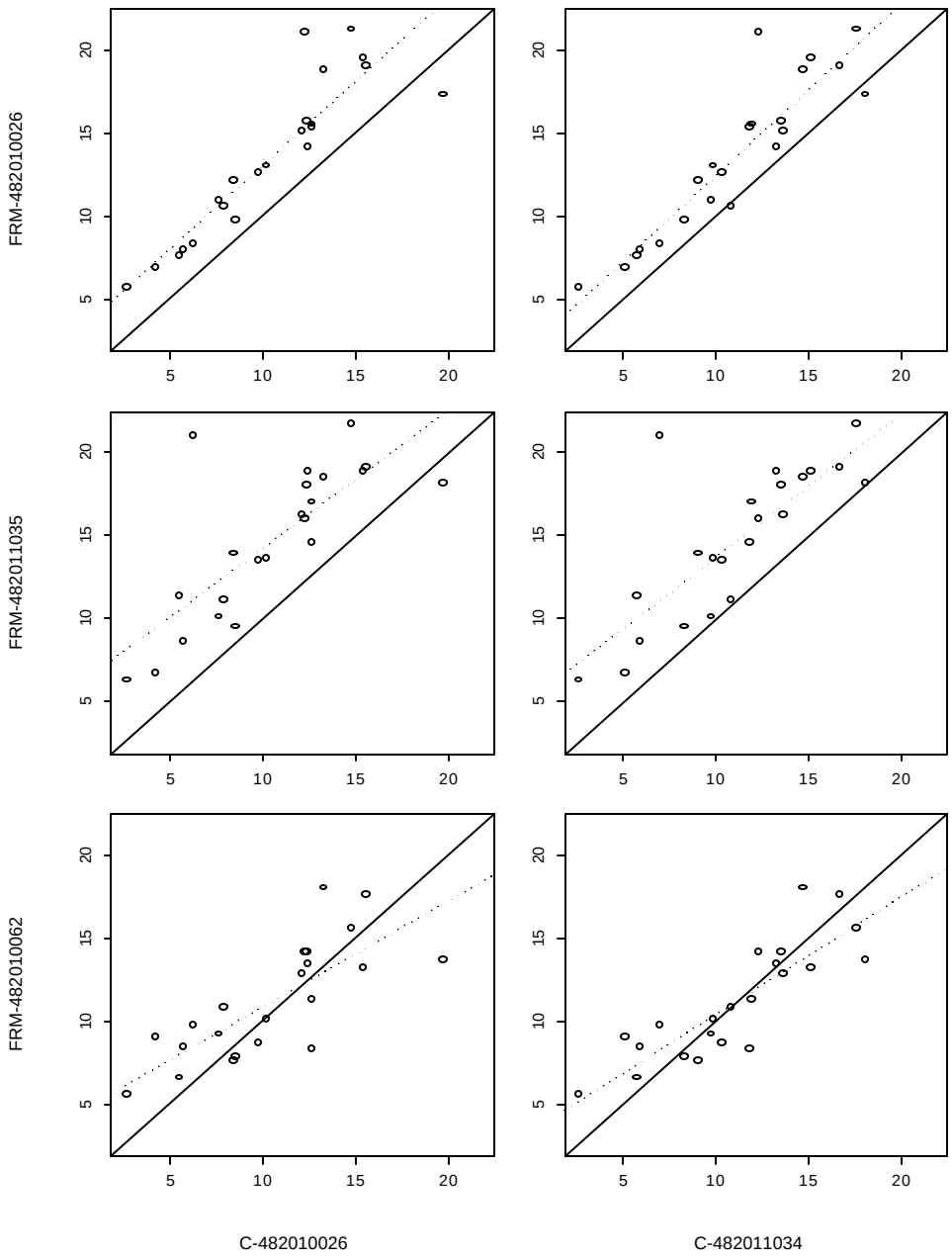

Figure B-33. Scatter plot of each of the three FRMs versus the two CMs. The solid line is the 45-degree line and the dashed line is the simple least squares regression line.

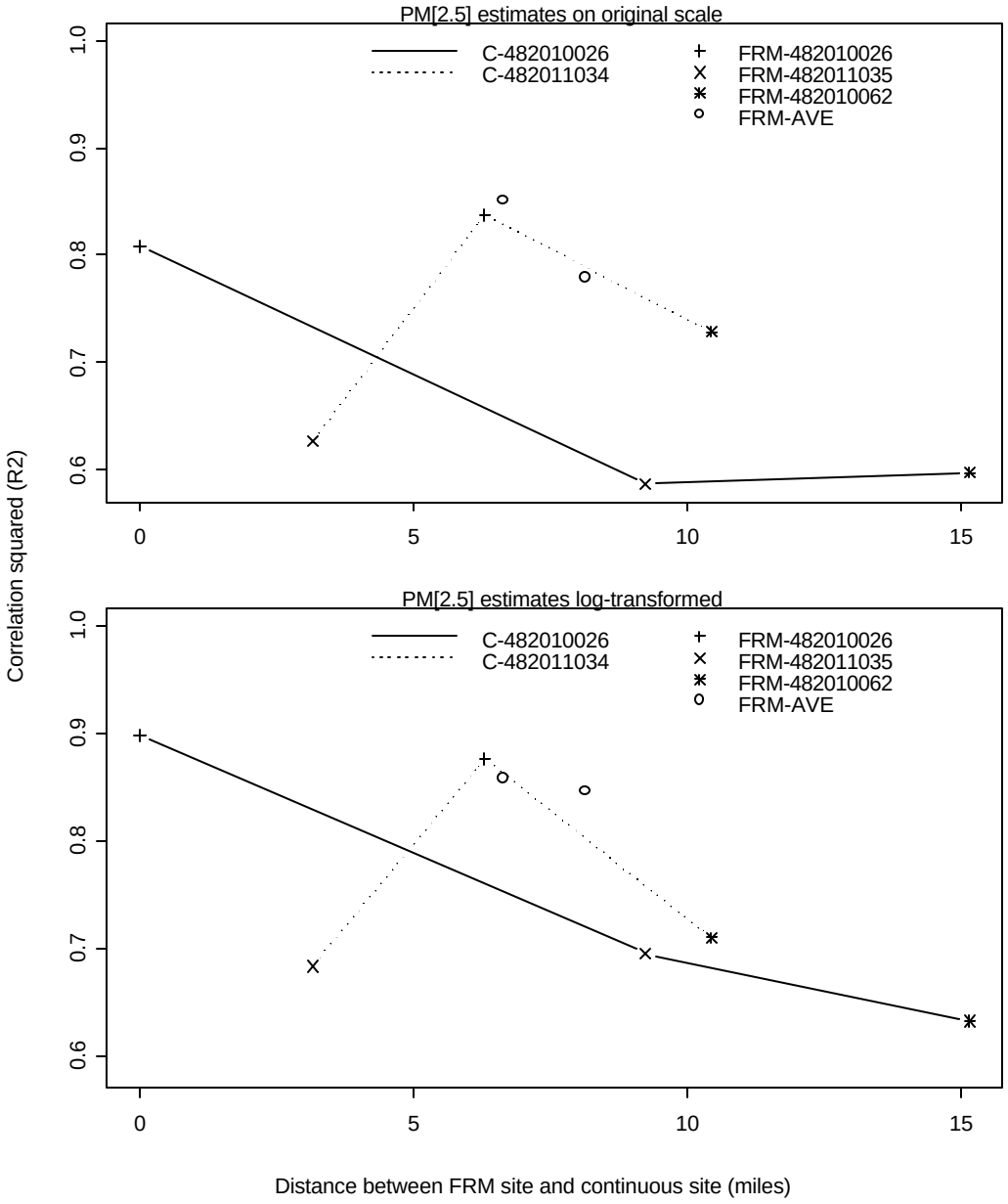

Figure B-34. **R^2 between different FRM monitors (different symbols) and CMs (different line types), plotted versus the distance between the sites. The two graphs correspond to $PM_{2.5}$ estimates on the original scale (top) and on the log-scale (bottom).**

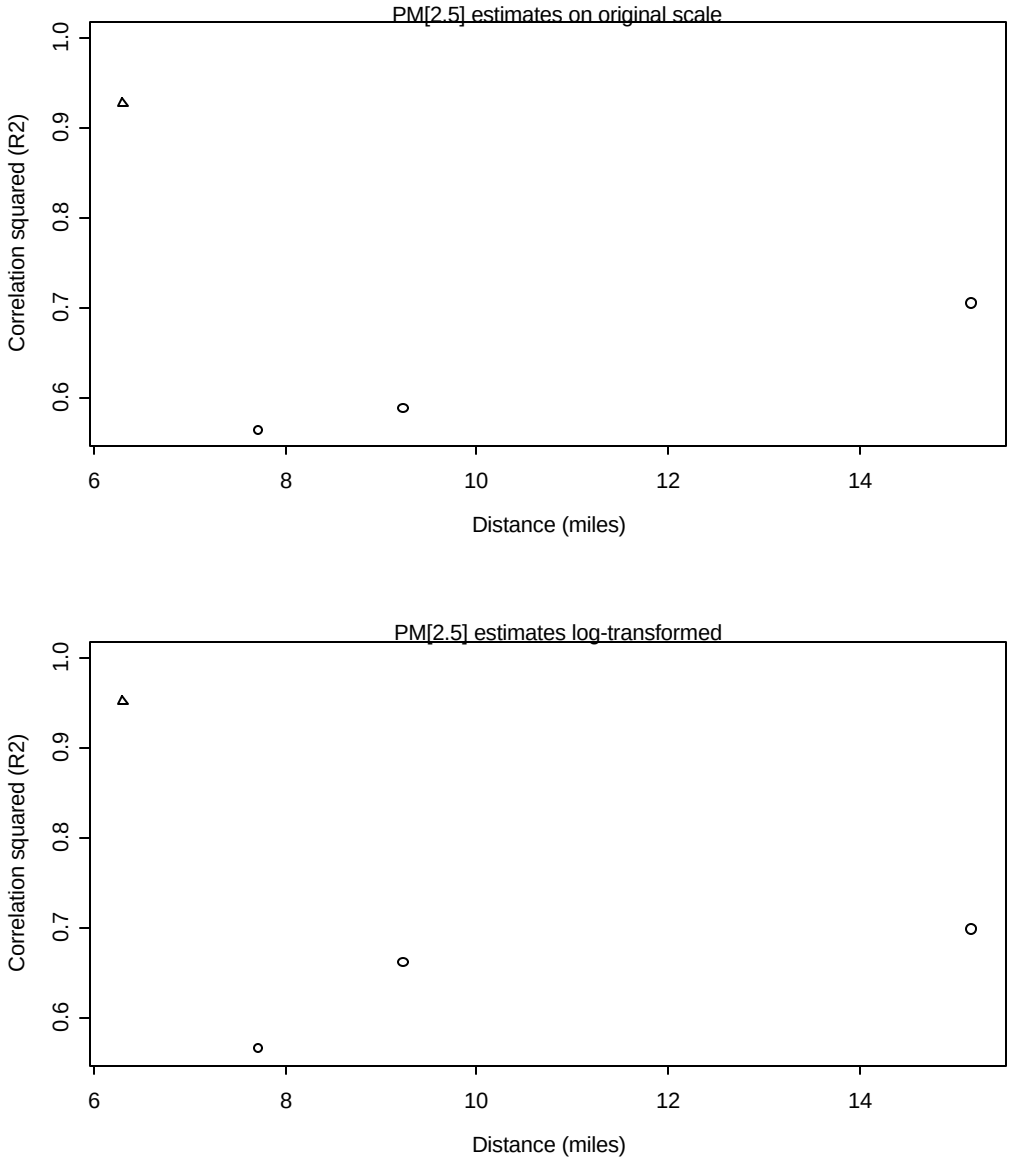

Figure B-35. **R^2 between the two CMs (one number shown as triangle in graphs) and between the three FRMs (three comparisons shown as circles in graphs), plotted versus the distance between the monitors. The two graphs correspond to $PM_{2.5}$ estimates on the original scale (top) and on the log-scale (bottom).**

TECHNICAL REPORT DATA

(Please read Instructions on reverse before completing)

1. REPORT NO.	2.	3. RECIPIENT'S ACCESSION NO.
EPA-454/R-01-002		

4. TITLE AND SUBTITLE
Data Quality Objectives (DQOs) and Model Development for Relating Federal Reference Method (FRM) and Continuous $PM_{2.5}$ Measurements to Report an Air Quality Index (AQI)

5. REPORT DATE
February, 2001

6. PERFORMING ORGANIZATION CODE

7. AUTHOR(S)

Shelly Eberly - U.S. EPA
Terence Fitz-Simons - U.S. EPA
Tim Hanley - U.S. EPA
Lewis Weinstock - Forsyth County NC, Environmental Affairs Department
Tom Tamanini - Hillsborough County FL, Environmental Protection Commission
Ginger Denniston - Texas Natural Resource Conservation Commission
Bryan Lambath - Texas Natural Resource Conservation Commission
Ed Michel - Texas Natural Resource Conservation Commission
Steve Bortnick - Battelle Memorial Institute

8. PERFORMING ORGANIZATION REPORT NO.

9. PERFORMING ORGANIZATION NAME AND ADDRESS

U.S. Environmental Protection Agency
Office of Air Quality Planning and Standards
Research Triangle Park, NC 27711

10. PROGRAM ELEMENT NO.

11. CONTRACT/GRANT NO.
68-D-98-030

12. SPONSORING AGENCY NAME AND ADDRESS
Director
Office of Air Quality Planning and Standards
Office of Air and Radiation
U.S. Environmental Protection Agency
Research Triangle Park, NC 27711

13. TYPE OF REPORT AND PERIOD COVERED
Final

14. SPONSORING AGENCY CODE
EPA/200/04

15. SUPPLEMENTARY NOTES
Prepared as a product of a Data Quality Objective (DQO) planning team.

16. ABSTRACT

All Metropolitan Statistical Areas (MSAs) with a population of 350,000 or greater are required to report daily air quality using the Air Quality Index (AQI) to the general public. According to Part 58 of 40 CFR, Appendix G, particulate matter measurements from non-Federal Reference Method (FRM) monitors may be used for the purpose of reporting the AQI if a linear relationship between these measurements and reference or equivalent method measurements can be established by statistical linear regression. This report provides guidance to MSA's for establishing a relationship between FRM and continuous $PM_{2.5}$ measurements. Chapter 2 of this report details the use of the EPA's Data Quality Objectives (DQOs) process to develop a statistical linear regression model relating FRM and continuous $PM_{2.5}$ measurements. Chapter 3 of this report offers step-by-step guidance to MSA's for developing a regression model relating FRM and continuous $PM_{2.5}$ measurements. Provided is a discussion of data issues likely to be encountered and methods to address them. Real-world examples are used for illustration, and are based on data from Davenport-Moline-Rock Island, IA-IL; Greensboro–Winston-Salem–High Point, NC; Salt Lake City-Ogden, UT; and Houston, TX.

17. KEY WORDS AND DOCUMENT ANALYSIS

a. DESCRIPTORS	b. IDENTIFIERS/OPEN ENDED TERMS	c. COSATI Field/Group
$PM_{2.5}$ Data Quality Objectives Air Quality Index	Air Pollution Measurement	

18. DISTRIBUTION STATEMENT	19. SECURITY CLASS *(Report)*	21. NO. OF PAGES
Release Unlimited	Unclassified	97
	20. SECURITY CLASS *(Page)*	22. PRICE
	Unclassified	

EPA Form 2220-1 (Rev. 4-77) PREVIOUS EDITION IS OBSOLETE